HAL open science

Sensitivity indices for multivariate outputs

Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnès Lagnoux

To cite this version:

Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnès Lagnoux. Sensitivity indices for multivariate outputs. 2013. hal-00800847v1

HAL Id: hal-00800847

https://inria.hal.science/hal-00800847v1

Preprint submitted on 14 Mar 2013 (v1), last revised 17 Apr 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sensitivity indices for multivariate outputs

Fabrice Gamboa [a], Alexandre Janon [b] Thierry Klein [a] Agnès Lagnoux [a]

[a] *Institut Mathématique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex.*

[b] *Laboratoire SAF, ISFA Université Lyon 1, 50 avenue Tony Garnier, 69007 Lyon*

Received *****; accepted after revision +++++

Presented by

Abstract

We define and study a generalization of Sobol sensitivity indices for the case of a vector output. *To cite this article: F. Gamboa, A. Janon, T. Klein, A. Lagnoux, C. R. Acad. Sci. Paris, Ser. xx xxx (2013).*

Résumé

Indices de sensibilité pour sorties multivariées. Nous définissons et étudions une généralisation des indices de Sobol pour des sorties vectorielles. *Pour citer cet article : F. Gamboa, A. Janon, T. Klein, A. Lagnoux, C. R. Acad. Sci. Paris, Ser. xx xxx (2013).*

1. Introduction

Many mathematical models encountered in applied sciences involve a large number of poorly-known parameters as inputs. It is important for the practitioner to assess the impact of this uncertainty on the model output. An aspect of this assessment is sensitivity analysis, which aims to identify the most sensitive parameters. In other words, parameters that have the largest influence on the output. In global stochastic sensitivity analysis, the input variables are assumed to be independent random variables. Their probability distributions account for the practitioner's belief about the input uncertainty. This turns the model output into a random variable.

When the output is scalar, using the so-called Hoeffding decomposition [4], its total variance can be split down into different partial variances. Each of these partial variances measures the uncertainty on the output induced by the corresponding input variable. By considering the ratio of each partial variance to the total variance, we obtain a measure of importance for each input variable called the *Sobol index* or

Email addresses: `fabrice.gamboa@math.univ-toulouse.fr` (Fabrice Gamboa), `alexandre.janon@imag.fr` (Alexandre Janon), `thierry.klein@math.univ-toulouse.fr` (Thierry Klein), `lagnoux@univ-tlse2.fr` (Agnès Lagnoux).

sensitivity index of the variable [3]; the most sensitive parameters can then be identified and ranked as the parameters with the largest Sobol indices.
Generalization of the Sobol index for multivariate (vector) outputs has been considered in [2] in an empirical way. In this note, we consider and study a new generalization of Sobol indices for vector outputs. These indices stem from an Hoeffding decomposition and satisfy natural invariance properties. In this note, we define the new sensitivity indices, examine some of their properties and show why they are natural. We also study a Monte-Carlo estimator of these indices, as in practice the exact values are not explicitly computable.

2. Definitions and Properties

2.1. *Preliminaries*

We denote by $X_1, \ldots, X_p$ the input random variables defined on some probability space $(\Omega, \mathbb{P})$, and by Y the output: $Y = f(X_1, \ldots, X_p)$, where $f : \mathbb{R}^p \to \mathbb{R}^k$ (p, k are integers). We suppose that $X_1, \ldots, X_p$ are independent, that $Y \in L^2(\Omega, \mathbb{R}^k)$, and that the covariance matrix of Y is positive definite.
For any non-empty r-subset $\mathbf{u}$ of $\{1, \ldots, p\}$, we set $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ and $X_{\sim\mathbf{u}} = (X_i, i \in \{1, \ldots, p\} \setminus \mathbf{u})$.

2.2. *Definition of the generalized Sobol indices*

We recall the Hoeffding decomposition of f [4]:

$$f(X_1, \ldots, X_p) = c + f_{\mathbf{u}}(X_{\mathbf{u}}) + f_{\sim\mathbf{u}}(X_{\sim\mathbf{u}}) + f_{\mathbf{u},\sim\mathbf{u}}(X_{\mathbf{u}}, X_{\sim\mathbf{u}}), \tag{1}$$

where $c \in \mathbb{R}^k$, $f_{\mathbf{u}} : \mathbb{R}^r \to \mathbb{R}^k$, $f_{\sim\mathbf{u}} : \mathbb{R}^{p-r} \to \mathbb{R}^k$ and $f_{\mathbf{u},\sim\mathbf{u}} : \mathbb{R}^p \to \mathbb{R}^k$ are given by:

$$c = \mathbb{E}(Y),\ f_{\mathbf{u}} = \mathbb{E}(Y|X_{\mathbf{u}}) - c,\ f_{\sim\mathbf{u}} = \mathbb{E}(Y|X_{\sim\mathbf{u}}) - c,\ f_{u,\sim\mathbf{u}} = Y - f_{\mathbf{u}} - f_{\sim\mathbf{u}} - c.$$

Taking the covariance matrices of both sides of (1) gives (thanks to L^2-orthogonality):

$$\Sigma = C_{\mathbf{u}} + C_{\sim\mathbf{u}} + C_{\mathbf{u},\sim\mathbf{u}}, \tag{2}$$

where Σ, $C_{\mathbf{u}}$, $C_{\sim\mathbf{u}}$ and $C_{\mathbf{u},\sim\mathbf{u}}$ are, respectively, the covariance matrices of Y, $f_{\mathbf{u}}(X_{\mathbf{u}})$, $f_{\sim\mathbf{u}}(X_{\sim\mathbf{u}})$ and $f_{\mathbf{u},\sim\mathbf{u}}(X_{\mathbf{u}}, X_{\sim\mathbf{u}})$.
For scalar outputs (ie., when $k = 1$), the covariance matrices are scalar (variances), and (2) is interpreted as the decomposition of the total variance of Y as a sum of the variance caused by the variation of the input factors X_i for $i \in \mathbf{u}$, the variance caused by the input factors not in $\mathbf{u}$, and the variance caused by the interactions of the factors in $\mathbf{u}$ and those not in $\mathbf{u}$. The (univariate) closed Sobol index $S^{\mathbf{u},\mathrm{Scal}}(f) = \frac{C_{\mathbf{u}}}{\Sigma}$ is then interpreted as the sensibility of Y to the inputs in $\mathbf{u}$. Due to noncommutativity of the matrix product, a direct generalization of this index is not straightforward.
We now go back to the general case. For any $k \times k$ matrix M, (2) can be projected on a scalar by multiplying by M and taking the trace:

$$\mathrm{Tr}(M\Sigma) = \mathrm{Tr}(MC_{\mathbf{u}}) + \mathrm{Tr}(MC_{\sim\mathbf{u}}) + \mathrm{Tr}(MC_{\mathbf{u},\sim\mathbf{u}}).$$

This equation is the natural scalarization of the matricial identity (2). This suggests to define, when $\mathrm{Tr}(M\Sigma) \neq 0$:

$$S^{\mathbf{u}}(M; f) = \frac{\mathrm{Tr}(MC_{\mathbf{u}})}{\mathrm{Tr}(M\Sigma)}$$

as the M-sensitivity measure (sensitivity index, or generalized Sobol index) of Y to the inputs in $\mathbf{u}$. We can also analogously define: $S^{\sim\mathbf{u}}(M; f) = \frac{\mathrm{Tr}(MC_{\sim\mathbf{u}})}{\mathrm{Tr}(M\Sigma)}$, $S^{\mathbf{u},\sim\mathbf{u}}(M; f) = \frac{\mathrm{Tr}(MC_{\mathbf{u},\sim\mathbf{u}})}{\mathrm{Tr}(M\Sigma)}$, which measures the

sensitivity to, respectively, the inputs not in $\mathbf{u}$, and to the interaction between inputs of $\mathbf{u}$ and inputs of $\{1,\dots,p\}\setminus\mathbf{u}$. The following lemma is obvious:

Lemma 2.1 *(i) The generalized sensitivity measures sum up to 1:*

$$S^{\mathbf{u}}(M;f)+S^{\sim\mathbf{u}}(M;f)+S^{\mathbf{u},\sim\mathbf{u}}(M;f)=1. \tag{3}$$

(ii) Left-composing f by a linear operator O of $\mathbb{R}^k$ changes the sensitivity measure according to:

$$S^{\mathbf{u}}(M;Of)=\frac{\mathrm{Tr}(MOC_{\mathbf{u}}O^t)}{\mathrm{Tr}(MO\Sigma O^t)}=\frac{\mathrm{Tr}(O^tMOC_{\mathbf{u}})}{\mathrm{Tr}(O^tMO\Sigma)}=S^{\mathbf{u}}(O^tMO;f). \tag{4}$$

(iii) For $k=1$, and for any $M\neq 0$, we have $S^{\mathbf{u}}(M;f)=S^{\mathbf{u},Scal}(f)$.

2.3. *The case $M=Id_k$*

We now consider the special case $M=\mathrm{Id}_k$ (the identity matrix of dimension k). We set $S_{\mathbf{u}}(f)=S^{\mathbf{u}}(\mathrm{Id}_k;f)$. The index $S_{\mathbf{u}}$ has the following properties:

Proposition 2.2 *Suppose that $Y\in L^2(\Omega,\mathbb{R}^k)$ and that Σ is positive-definite:*

(i) $0\le S_{\mathbf{u}}(f)\le 1$;

(ii) $S_{\mathbf{u}}(f)$ is invariant by left-composition of f by any isometry of $\mathbb{R}^k$, i.e.

$$\forall O\ k\times k \text{ matrix s.t. } O^tO=Id_k,\ \ S_{\mathbf{u}}(Of)=S_{\mathbf{u}}(f);$$

(iii) $S_{\mathbf{u}}(f)$ is invariant by left-composition of f by any nonzero homothety of $\mathbb{R}^k$.

Proof: Point (i): positivity is clear, as $C_{\mathbf{u}}$ and Σ are positive; $S_{\mathbf{u}}(f)\le 1$ follows from positivity and (3). For (ii), we use (4). Point (iii) is immediate. □

The properties in the Proposition above are natural requirements for a sensitivity measure. In the next section, we will show that these requirements can be fulfilled by $S^{\mathbf{u}}(M;\cdot)$ iff $M=\lambda\mathrm{Id}_k$ for $\lambda\in\mathbb{R}^*$. Hence the only "canonical" sensitivity measure is $S_{\mathbf{u}}$.

3. $M=\mathrm{Id}_k$ is the only good choice

The following Proposition can be seen as a kind of converse of Proposition 2.2.

Proposition 3.1 *Let M be a square $k\times k$ matrix such that:*

(i) M does not depend on f, p nor $\mathbf{u}$;

(ii) $\mathrm{Tr}(MV)\neq 0$ for any symmetric positive-definite matrix V;

(iii) $S^{\mathbf{u}}(M;f)$ is invariant by left-composition of f by any isometry of $\mathbb{R}^k$.

Then $S^{\mathbf{u}}(M;\cdot)=S_{\mathbf{u}}$.

Proof: Let M as in the Proposition. We can write $M=M_{Sym}+M_{Antisym}$ where $M_{Sym}^t=M_{Sym}$ and $M_{Antisym}^t=-M_{Antisym}$. Since, for any matrix V, we have $\mathrm{Tr}(M_{Antisym}V)=0$, we have $S^{\mathbf{u}}(M;f)=S^{\mathbf{u}}(M_{Sym};f)$ and we can assume, without loss of generality, that M is symmetric.

We diagonalize M in an orthonormed basis: $M=PDP^t$, where $P^tP=\mathrm{Id}_k$ and D diagonal. We have:

$$S^{\mathbf{u}}(M;f)=\frac{\mathrm{Tr}(PDP^tC_{\mathbf{u}})}{\mathrm{Tr}(PDP^t\Sigma)}=\frac{\mathrm{Tr}(DP^tC_{\mathbf{u}}P)}{\mathrm{Tr}(DP^t\Sigma P)}=S^{\mathbf{u}}(D;P^tf).$$

This (thanks to hypothesis (i)) shows that M can in fact be assumed diagonal.

Now we want to show that $M = \lambda \mathrm{Id}_k$ for some $\lambda \in \mathbb{R}^*$. Suppose, by contradiction, that M has two different diagonal coefficients $\lambda_1 \neq \lambda_2$. It is clearly sufficient to consider the case $k = 2$. Choose $f = \mathrm{Id}_2$ (hence, $p = 2$), and $\mathbf{u} = \{1\}$ (these choices are valid by (i) and (ii)). We have $\Sigma = \mathrm{Id}_2$, and $C_{\mathbf{u}} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, hence on the one hand $S^{\mathbf{u}}(M; f) = \frac{\lambda_1}{\lambda_1+\lambda_2}$. On the other hand, let O be the isometry which exchanges the two vectors of the canonical basis of $\mathbb{R}^2$. We have $S^{\mathbf{u}}(M; Of) = \frac{\lambda_2}{\lambda_1+\lambda_2}$, hence (iii) is contradicted if $\lambda_1 \neq \lambda_2$. The case $\lambda = 0$ is forbidden by (ii). Finally, it is easy to check that, for any $\lambda \in \mathbb{R}^*$, $S^{\mathbf{u}}(\lambda \mathrm{Id}_k; \cdot) = S^{\mathbf{u}}(\mathrm{Id}_k; \cdot) = S_{\mathbf{u}}$. $\square$

4. Estimation of $S^{\mathbf{u}}$

In general, the covariance matrices $C_{\mathbf{u}}$ and Σ are not analytically available. In the scalar case ($k = 1$), it is customary to estimate $S^{\mathbf{u},\mathrm{Scal}}(f)$ by using a Monte-Carlo pick-freeze method [3, 1], which uses a finite sample of evaluations of f. In this Section, we propose a pick-freeze estimator for the vector case which generalizes the T_N estimator studied in [1]. We set: $Y^{\mathbf{u}} = f(X_{\mathbf{u}}, X'_{\sim\mathbf{u}})$ where $X'_{\sim\mathbf{u}}$ is an independent copy of $X_{\sim\mathbf{u}}$. Let N be an integer. We take N independent copies $Y_1, \ldots, Y_N$ (resp. $Y_1^{\mathbf{u}}, \ldots, Y_N^{\mathbf{u}}$) of Y (resp. $Y^{\mathbf{u}}$). For $l = 1, \ldots, k$, and $i = 1, \ldots, N$, we also denote by $Y_{i,l}$ (resp. $Y_{i,l}^{\mathbf{u}}$) the lth component of Y_i (resp. $Y_i^{\mathbf{u}}$). We define the following estimator of $S_{\mathbf{u}}(f)$:

$$S_{\mathbf{u},N} = \frac{\sum_{l=1}^{k} \left(\sum_{i=1}^{N} Y_{i,l} Y_{i,l}^{\mathbf{u}} - \frac{1}{N} \left(\sum_{i=1}^{N} \frac{Y_{i,l} + Y_{i,l}^{\mathbf{u}}}{2} \right)^2 \right)}{\sum_{l=1}^{k} \left(\sum_{i=1}^{N} \frac{Y_{i,l}^2 + (Y_{i,l}^{\mathbf{u}})^2}{2} - \frac{1}{N} \left(\sum_{i=1}^{N} \frac{Y_{i,l} + Y_{i,l}^{\mathbf{u}}}{2} \right)^2 \right)}.$$

The following Proposition can be proved in a way similar to the one used to prove Proposition 2.2 and Proposition 2.5 of [1].

Proposition 4.1 *Suppose $Y \in L^4(\Omega, \mathbb{R}^k)$, and that Σ is positive-definite. Then:*

(i) $(S_{\mathbf{u},N}(f))_N$ is asymptotically normal: there exists $\sigma = \sigma(f)$ so that $\sqrt{N}(S_{\mathbf{u},N}(f) - S_{\mathbf{u}}(f))$ converges (for $N \to +\infty$) in distribution to a centered Gaussian distribution with variance σ^2.

(ii) $(S_{\mathbf{u},N}(f))_N$ is asymptotically efficient for estimating $S_{\mathbf{u}}(f)$ among regular estimator sequences that are function of exchangeable pairs $(Y, Y^{\mathbf{u}})$.

Acknowledgements. This work has been partially supported by the French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA nANR-09-COSI-015).

References

[1] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. Preprint available at `http://hal.inria.fr/hal-00665048/en`, 2012.

[2] Matieyendou Lamboni, Hervé Monod, and David Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4):450–459, 2011.

[3] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995), 1993.

[4] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.