Normalité asymptotique et efficacité dans l'estimation des indices de Sobol

Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, Clémentine Prieur

Normalité asymptotique et efficacité dans l'estimation des indices de Sobol

Alexandre Janon[1], Thierry Klein[2], Agnès Lagnoux[2], Maëlle Nodet[1] & Clémentine Prieur[1]

[1] *Laboratoire Jean Kuntzman, Université Joseph Fourier (Grenoble 1), INRIA/MOISE, 51 rue des Mathématiques, BP 53, 38041 Grenoble cedex 9, France*

[2] *Laboratoire de Statistique et Probabilités, Institut de Mathématiques, Université Paul Sabatier (Toulouse 3) 31062 Toulouse Cedex 9, France*

alexandre.janon@imag.fr, thierry.klein@math.univ-toulouse.fr, agnes.lagnoux@math.univ-toulouse.fr, maelle.nodet@inria.fr, clementine.prieur@imag.fr

Résumé. De nombreux modèles mathématiques font intervenir plusieurs paramètres qui ne sont pas tous connus précisément. L'analyse de sensibilité globale se propose de sélectionner les paramètres d'entrée dont l'incertitude a le plus d'impact sur la variabilité d'une quantité d'intérêt, sortie du modèle. Un des outils statistiques pour quantifier l'influence de chacune des entrées sur la sortie est l'indice de sensibilité de Sobol. Nous considérons l'estimation statistique de cet indice à l'aide d'un nombre fini d'échantillons de sorties du modèle : nous présentons deux estimateurs de cet indice et énonçons un théorème central limite pour chacun d'eux. Nous démontrons que l'un de ces deux estimateurs est optimal en terme de variance asymptotique. Nous généralisons également nos résultats au cas où la vraie sortie du modèle n'est pas observée, mais où seule une version dégradée (bruitée) de la sortie est disponible.

Mots-clés. Analyse de sensibilité globale, indice de Sobol, méthode de Monte Carlo, efficacité asymptotique, surfaces de réponse, métamodèle.

Abstract. Many mathematical models involve input parameters, which are not precisely known. Global sensitivity analysis aims to identify the parameters whose uncertainty has the largest impact on the variability of a quantity of interest (output of the model). One of the statistical tools used to quantify the influence of each input variable on the output is the Sobol sensitivity index. We consider the statistical estimation of this index from a finite sample of model outputs : we present two estimators and state a central limit theorem for each. We show that one of these estimators has an optimal asymptotic variance. We also generalize our results to the case where the true output is not observable, and is replaced by a noisy version.

Keywords. Global sensitivity analysis, Sobol index, Monte Carlo method, asymptotic efficiency, response surface method, metamodel.

1 Définition et estimation des indices de Sobol

1.1 Motivation et définition des indices de sensibilité de Sobol

La plupart des modèles étudiés en mathématiques appliquées font intervenir de nombreux paramètres qui ne sont pas tous connus avec précision. Il est alors important, pour une utilisation correcte du modèle, de pouvoir quantifier l'impact de l'incertitude attachée aux paramètres d'entrée sur une quantité d'intérêt en sortie du modèle. Dans ce contexte, l'analyse de sensibilité (Saltelli et al. (2004)) cherche à identifier les paramètres d'entrée les plus influents sur la sortie, et à quantifier l'importance prise par chacune des variables d'entrée sur la sortie. L'utilisation des indices de Sobol est une méthode répandue en analyse de sensibilité, qui repose sur l'attribution d'une loi de probabilité (supposée connue) aux variables d'entrée afin de modéliser l'incertitude sur chaque paramètre d'entrée.

Plus précisément, notons $X_1, \dots, X_p$ les variables d'entrées et $f : \mathbb{R}^p \to \mathbb{R}$ la fonction reliant les variables d'entrées à la sortie (quantité d'intérêt). On suppose que les variables aléatoires $X_1, \dots, X_p$ sont indépendantes et que $Y = f(X_1, \dots, X_p)$ est une variable aléatoire L^2 telle que $\operatorname{Var} Y \neq 0$. On définit pour $i = 1, \dots, p$ l'indice de Sobol de Y par rapport à X_i par la formule suivante :

$$S_i = \frac{\operatorname{Var} \mathbb{E}(Y|X_i)}{\operatorname{Var} Y}$$

Cet indice, compris entre 0 et 1, quantifie la variance de Y due à la variabilité de X_i seule (au numérateur) relativement à la variance totale de Y (au dénominateur).

Dans la majorité des cas pratiques, la fonction f n'est pas connue explicitement à l'aide d'une formule analytique simple, ce qui empêche le calcul direct de S_i à l'aide de sa définition. Cependant, la sortie $f(X_1, \dots, X_p)$ peut être évaluée numériquement, pour une valeur donnée de $(X_1, \dots, X_p)$, à l'aide d'un code informatique. Il est alors nécessaire de considérer l'*estimation*, au sens statistique du terme, de S_i à partir d'un *échantillon* fini d'évaluations de f.

1.2 Estimation des indices de sensibilité

Une technique classique d'estimation de S_i consiste, pour $N \in \mathbb{N}$, à considérer les $2N$ évaluations de f suivantes : pour $k = 1, \dots, N$,

$$Y_k = f\left(X_{1,k}, X_{2,k}, \dots, X_{p,k}\right), \quad Y'_k = f\left(X'_{1,k}, X'_{2,k}, \dots, X'_{i-1,k}, X_{i,k}, X'_{i+1,k}, \dots, X'_{p,k}\right),$$

où $\{(X_{1,k}, \dots, X_{p,k})\}_{k=1,\dots,N}$ et $\left\{\left(X'_{1,k}, \dots, X'_{p,k}\right)\right\}_{k=1,\dots,N}$ sont deux échantillons iid de la loi de $(X_1, \dots, X_p)$, puis à estimer l'indice S_i par :

$$\widehat{S}_{i,N} = \frac{\frac{1}{N}\sum_{k=1}^{N} Y_k Y'_k - \left(\frac{1}{N}\sum_{k=1}^{N} Y_k\right)\left(\frac{1}{N}\sum_{k=1}^{N} Y'_k\right)}{\frac{1}{N}\sum_{k=1}^{N} Y_k^2 - \left(\frac{1}{N}\sum_{k=1}^{N} Y_k\right)^2}.$$

Cet estimateur est considéré dans Homma et Saltelli (1996).

Dans Janon et al. (2012) nous proposons d'utiliser l'estimateur

$$\widehat{T}_{i,N} = \frac{\frac{1}{N}\sum_{k=1}^{N} Y_k Y'_k - \left(\frac{1}{N}\sum_{k=1}^{N}\left[\frac{Y_k+Y'_k}{2}\right]\right)^2}{\frac{1}{N}\sum_{k=1}^{N}\left[\frac{Y_k^2+(Y'_k)^2}{2}\right] - \left(\frac{1}{N}\sum_{k=1}^{N}\left[\frac{Y_k+Y'_k}{2}\right]\right)^2}$$

2 Propriétés asymptotiques des estimateurs

Dans la suite, nous énonçons les propriétés asymptotiques (quand $N \to +\infty$) des estimateurs $\widehat{S}_{i,N}$ et $\widehat{T}_{i,N}$ démontrées dans Janon et al. (2012).

2.1 Utilisation du "vrai" modèle

Supposons que $Y \in L^4$. Alors les estimateurs $\widehat{S}_{i,N}$ et $\widehat{T}_{i,N}$ sont asymptotiquement normaux, c'est-à-dire que nous avons les théorèmes limites centraux suivants

$$\sqrt{N}\left(\widehat{S}_{i,N} - S_i\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma_S^2\right),$$

et :

$$\sqrt{N}\left(\widehat{T}_{i,N} - S_i\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma_T^2\right)$$

où $\mathcal{N}(0,\sigma^2)$ désigne la loi normale centrée de variance σ^2 et où les variances asymptotiques sont données par

$$\sigma_S^2 = \frac{\text{Var}\ \left((Y - \mathbb{E}(Y))\left[(Y' - \mathbb{E}(Y)) - S_i(Y - \mathbb{E}(Y))\right]\right)}{(\text{Var}\ Y)^2}$$

$$\sigma_T^2 = \frac{\text{Var}\ \left((Y - \mathbb{E}(Y))(Y' - \mathbb{E}(Y)) - S_i/2\left((Y - \mathbb{E}(Y))^2 + (Y' - \mathbb{E}(Y))^2\right)\right)}{(\text{Var}\ Y)^2}.$$

Ces variances limites peuvent être estimées, ce qui permet de construire des intervalles de confiance asymptotiques pour S_i, de longueur proportionnelle à $\sigma_S/\sqrt{N}$ (resp. $\sigma_T/\sqrt{N}$) si l'estimateur $\widehat{S}_{i,N}$ (resp. $\widehat{T}_{i,N}$) est utilisé.

Dans ce contexte, l'estimateur le plus efficace est celui donnant l'intervalle de confiance asymptotique le moins étendu, c'est-à-dire celui ayant la plus petite variance asymptotique. Nous justifions l'introduction de l'estimateur $\widehat{T}_{i,N}$ par le fait que sa variance asymptotique est toujours inférieure ou égale à celle de $\widehat{S}_{i,N}$, les seuls cas d'égalité étant les cas $S_i = 0$ ou $S_i = 1$.

Il est même possible de démontrer que la variance asymptotique de $\widehat{T}_{i,N}$ est inférieure ou égale à la variance de *tout* estimateur régulier de S_i construit à partir des observations de $(Y_k, Y'_k)_k$. Cette propriété se nomme *efficacité asymptotique* (van der Vaart (2000), chapitres 8 et 25) et généralise la notion d'estimateur non biaisé de variance minimale, basée sur l'inégalité de Cramér-Rao.

2.2 Utilisation d'un métamodèle

Comme dit dans l'introduction, l'évaluation de la fonction f nécessite généralement l'appel à un code d'approximation numérique ; chaque appel à ce code peut être coûteux en temps de calcul. Comme l'évaluation des estimateurs $\widehat{S}_{i,N}$ et $\widehat{T}_{i,N}$ requiert un nombre important d'évaluations de la fonction f, il est souvent nécessaire de remplacer les appels à f par des appels à une fonction $\widetilde{f}$ qui approche f tout en étant beaucoup moins coûteuse en temps de calcul (la fonction $\widetilde{f}$ est appelée métamodèle, ou surface de réponse, pour f).

Dans ce cadre, on obtient des estimateurs $\widetilde{S}_{i,N}$ et $\widetilde{T}_{i,N}$, qui estiment l'indice de Sobol du métamodèle : $\widetilde{S}_i = \frac{\text{Var } \mathbb{E}(\widetilde{f}(X_1,\ldots,X_p)|X_i)}{\text{Var } \widetilde{f}(X_1,\ldots,X_p)}$. Ces estimateurs ne sont cependant jamais consistants pour S_i, à moins bien sûr que $S_i = \widetilde{S}_i$.

Afin d'obtenir des résultats asymptotiques sur $\widetilde{S}_{i,N}$ et $\widetilde{T}_{i,N}$ relativement à l'estimation de S_i, il est nécessaire, et naturel, de supposer que l'erreur de métamodèle $f - \widetilde{f}$ dépend de N et que son impact sur l'estimation de S_i, en un sens à préciser, tend vers 0 lorsque N tend vers $+\infty$. Notons que le métamodèle dépend alors de N : $\widetilde{f} = \widetilde{f}_N$.

Nous démontrons alors que, sous des hypothèses techniques de domination, la convergence en loi de $\sqrt{N}(\widetilde{S}_{i,N} - S_i)$ (resp. $\sqrt{N}(\widetilde{T}_{i,N} - S_i)$) vers $\mathcal{N}(0, \sigma_S^2)$ (resp. $\mathcal{N}(0, \sigma_T^2)$) a lieu si et seulement si Var $(f - \widetilde{f}_N) = o(1/N)$.

Sous d'autres hypothèses techniques, ainsi que sous l'hypothèse $\mathbb{E}(f - \widetilde{f}_N) = o(1/\sqrt{N})$, nous démontrons également que $\widetilde{T}_{i,N}$ est asymptotiquement efficace pour S_i.

3 Résultats numériques

Dans cette section, nous illustrons nos résultats de normalité asymptotique sur le modèle de test suivant (fonction d'Ishigami) :

$$f(X_1, X_2, X_3) = \sin X_1 + 7 \sin^2 X_2 + 0.1 X_3^4 \sin X_1,$$

où $(X_j)_{j=1,2,3}$ sont des variables iid dans $[-\pi; \pi]$. Les valeurs des indices de sensibilité pour cette fonction sont connues analytiquement : $S_1 = 0.3139, S_2 = 0.4424, S_3 = 0$. Ceci permet de calculer le *coverage* empirique des intervalles de confiance construits à l'aide des résultats de normalité asymptotique, c'est-à-dire la proportion, sur mille intervalles de confiance calculés, d'intervalles contenant la vraie valeur de l'indice. Cette proportion doit être théoriquement proche du niveau choisi pour l'intervalle de confiance, fixé pour toutes nos simulations à $0,95$.

3.1 Observation du "vrai" modèle

La Figure 1 montre le coverage empirique de l'intervalle de confiance construit sur l'estimateur $\widehat{S}_{i,N}$, pour l'indice relatif à chacune des variables d'entrée X_1, X_2 et X_3. On

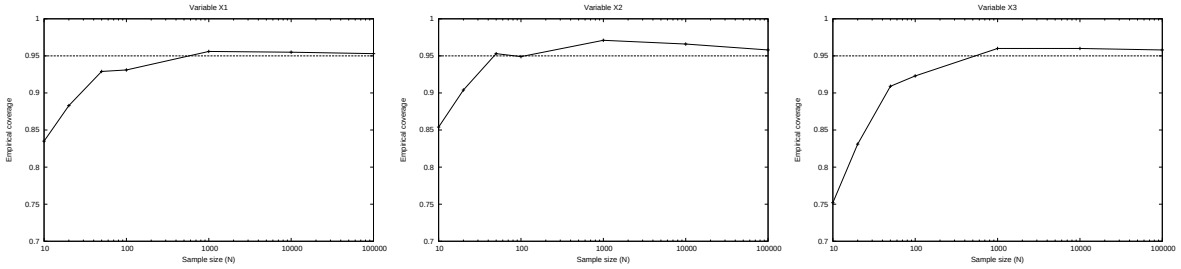

FIGURE 1 – Coverages empiriques des intervalles de confiance asymptotiques pour S_1 (gauche), S_2 (centre) et S_3 (droite), en fonction de la taille de l'échantillon. L'estimateur $\widehat{S}_{i,N}$ est utilisé.

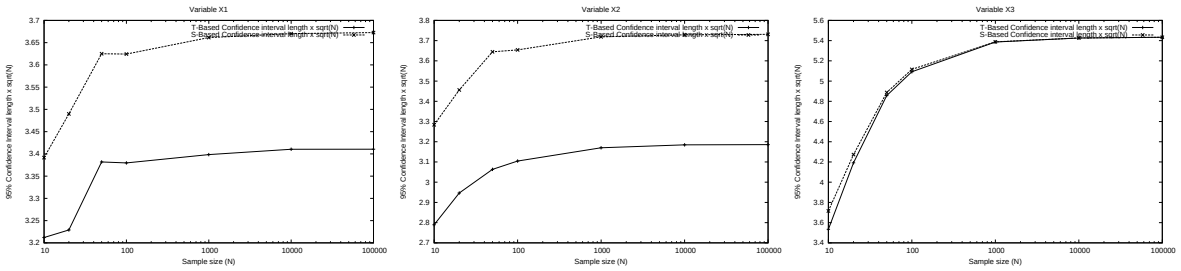

FIGURE 2 – Longueurs (renormalisées par $\sqrt{N}$) des intervalles de confiances à 95% pour S_1 (gauche), S_2 (centre), S_3 (droite). Traits pleins : utilisation de $\widehat{T}_{i,N}$, traits pointillés : utilisation de $\widehat{S}_{i,N}$.

observe ainsi la convergence vers le coverage théorique (marqué en lignes pointillées). La même convergence s'observe pour $\widehat{T}_{i,N}$.

Dans la Figure 2, nous comparons la taille (renormalisée d'un facteur $\sqrt{N}$) des intervalles de confiance produits grâce au théorème central limite sur $\widehat{S}_{i,N}$ (traits pointillés) et $\widehat{T}_{i,N}$ (traits pleins). Nous observons que l'intervalle construit à partir de $\widehat{T}_{i,N}$ est toujours strictement plus petit que celui construit à partir de $\widehat{S}_{i,N}$, sauf pour X_3 où les deux intervalles ont sensiblement même longueurs. Ceci s'accorde avec notre résultat d'optimalité pour $\widehat{T}_{i,N}$, et notre cas d'égalité, puisque $S_3 = 0$.

3.2 Observation d'un métamodèle RKHS

Nous considérons maintenant le cas où seul un métamodèle $\widetilde{f}_N$ approchant f est observé. Ce métamodèle est construit par interpolation dans un RKHS (*reproducing kernel Hilbert space*). Cette approche de métamodélisation, revue dans Scheuerer et al. (2011), et équivalente au Krigeage, utilise un échantillon d'apprentissage (un échantillon d'évaluations

a	N	n	Coverage pour S_1	Cov. pour S_2	Cov. pour S_3
.4	3000	33	0.1	0	0.7
.4	10000	51	0.28	0.18	0.78
.4	20000	77	0.28	0.1	0.59
.6	3000	111	0.79	0.37	0.9
.6	10000	169	0.92	0.82	0.94
.6	20000	210	0.93	0.85	0.95
.7	3000	177	0.93	0.88	0.93
.7	6000	226	0.94	0.93	0.97

FIGURE 3 – Coverages empiriques des intervalles de confiance asymptotiques estimés à partir de l'observation d'un métamodèle de f.

de f) de taille n et est donc intéressante en terme de temps de calcul si $n < 2N$. Bien sûr, plus n est grand, plus le métamodèle est coûteux mais plus l'erreur $f - \widetilde{f}_N$ est faible. L'analyse d'erreur de l'interpolation par RKHS, détaillée dans Janon et al. (2012) montre que si n et N sont reliés par $n = (a \ln N)^3$, notre condition de normalité asymptotique $\mathrm{Var}\,(f - \widetilde{f}_N) = o(1/N)$ a lieu pour $a > 0.52$.

Le tableau en Figure 3 montre que cette valeur critique pour a (valeur déterminant la vitesse de croissance de n en fonction de N) est déterminante pour la normalité asymptotique de $\widetilde{S}_{i,N}$, puisque le coverage théorique (0,95) est atteint seulement lorsque a dépasse cette valeur critique.

Remerciement. Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (ANR) au travers du programme COSINUS (projet COSTA-BRAVA n°ANR-09-COSI-015).

Bibliographie

[1] Homma, T. et Saltelli A. (1996), Importance measures in global sensitivity analysis of nonlinear models, *Reliability Engineering & System Safety*, 1–17.
[2] Janon, A., Klein, T., Lagnoux, A., Nodet, M. et Prieur, C. (2012), Asymptotic normality and efficiency of two Sobol index estimators, `http://hal.inria.fr/hal-00665048`.
[3] Saltelli, A., Tarantola, S., Campolongo, F. et Ratto, M. (2004), Sensitivity analysis in practice : a guide to assessing scientific models, *Wiley, New-York*.
[4] Scheuerer, M., Schaback, R. et Schlather, M. (2011), Interpolation of Spatial Data – A Stochastic or a Deterministic Problem ?
[5] Van der Vaart, A.W. (2000), Asymptotic Statistics, *Cambridge University Press*.