Plans emboîtés pour l'estimation itérative des indices de Sobol' par méthode répliquée

Laurent Gilquin [1] & Clémentine Prieur [2] & Elise Arnaud [2]

[1] *Inria Grenoble - Rhône-Alpes, Inovallée, 655 avenue de l'Europe, 38330 Montbonnot, laurent.gilquin@inria.fr*

[2] *LJK, Université de Grenoble, 51 rue des Mathématiques Campus de Saint Martin d'Hères, 38041 Grenoble cedex 09, prenom.nom@imag.fr*

Résumé. Ce travail s'intéresse à l'estimation d'indices de Sobol' d'ordre un et deux pour l'analyse de sensibilité. Dans ce cadre, l'utilisation de la méthode répliquée permet d'assurer un nombre d'appels considérablement réduit par rapport aux méthodes classiques. L'objectif de cette étude est de proposer une approche itérative par plans répliqués pour estimer les indices de sensibilité. L'élément clé est la construction de plans emboîtés. Nous proposons ici une adaptation de la méthode répliquée par l'utilisation d'un plan emboîté pour estimer les indices de Sobol' d'ordre un ou d'ordre deux globaux. Pour l'estimation des indices d'ordre un, nous exploitons un plan particulier ayant déjà été introduit dans la littérature. Pour l'estimation des indices d'ordre deux globaux, la méthode repose sur des tableaux orthogonaux. Nous présentons donc deux approches pour construire un tableau orthogonal de force deux emboîté. La première méthode est stochastique et repose sur des résultats de théorie des graphes. L'idée de la méthode est de combler itérativement les zones lacunaires de l'espace des paramètres d'entrée. La deuxième méthode consiste à construire un tableau orthogonal de force deux d'index supérieur à un, puis à rééchantillonner à l'intérieur de chaque cellule par une loi uniforme. Nous conduisons des tests numériques sur des fonctions classiques afin de comparer les indices d'ordre un et d'ordre deux globaux obtenus par chacune des deux méthodes à ceux obtenus par une méthode standard (non emboîté).

Mots-clés. analyse de sensibilité, indice de Sobol' groupé, plan emboîté, recouvrement de l'espace, tableau orthogonal

Abstract. This work deals with the estimation of first-order and second-order Sobol' indices used in sensitivity analysis. Compared to more classical methods, the replication method allows to considerably reduce the number of runs. The aim of this study is to propose an iterative approach, using replicated designs, to estimate the sensitivity indices. The key feature is the construction of nested designs. We propose here an adaptation of the replication method to use nested designs in the aim of estimate first-order and closed second-order Sobol' indices. For the estimation of first-order indices, we exploit a specific design already introduced in the literature. For the estimation of closed second-order Sobol' indices, the method relies on orthogonal arrays. We present two approaches to

construct a nested orthogonal array of strength two. The former is stochastic and relies on results from graph theory. It aims at iteratively fill the empty zones of the input space. The latter consists in constructing an orthogonal array of strength two of index greater than one; then to re-sample inside each cell with a uniform distribution. We conduct numerical experiments on classical functions to compare the estimated first- and closed second-order Sobol' indices of each of the two approaches to those obtained with a more standard one (without a nested design).

Keywords. sensitivity analysis, grouped Sobol' index, nested design, space-filling, orthogonal array

1 Contexte

Dans beaucoup de modèles mathématiques, les paramètres d'entrée sont souvent source d'incertitudes et peuvent avoir des effets significatifs sur la sortie du modèle. Il est important de mesurer ces effets pour les utilisateurs du modèle. L'analyse de sensibilité globale est une pratique commune pour identifier les paramètres d'entrée influents et détecter les interactions potentielles entre eux. Parmi le large panel de méthodes disponibles pour effectuer une telle analyse, la méthode basée sur la variance introduite par Sobol' (1993) permet de calculer des indices de sensibilité appelés indices de Sobol'.

ANOVA fonctionnelle et indices de Sobol'

Afin d'introduire ces indices, considérons un modèle représenté par une fonction f, un vecteur aléatoire $X = (X_1, \dots, X_d)$ de paramètres d'entrée et $Y = f(X)$ une sortie du modèle. Soit $P_X = P_{X_1} \otimes \dots \otimes P_{X_d}$ la distribution de X, nous supposons que $f \in \mathbb{L}^2(P_X)$. f admet une décomposition unique en composantes de dimensions croissantes (décomposition fonctionnelle ANOVA) :

$$f(X) = f_0 + \sum_i f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \dots + f_{1\dots d}(X_1, \dots, X_d) \ ,$$

où chaque composante vérifie :

$$\int f_{i_1\dots i_s}(x_{i_1}, \dots, x_{i_s}) dP_{X_{i_k}}(x_{i_k}) = 0, \ \forall k \in \{1, \dots, s\}, \forall i_1, \dots, i_s \in \{1, \dots, d\}.$$

La décomposition fonctionnelle peut être utilisée pour mesurer la sensibilité globale de la sortie Y par rapport au paramètre d'entrée X_i. Soit $I \subseteq \{1, \dots, d\}$, par orthogonalité on obtient :

$$\mathrm{Var}[Y] = \sum_i V_i + \sum_{i<j} V_{ij} + \dots + V_{1,\dots,d} \ \text{ où } \ V_I = \mathrm{Var}[f_I(X_I)] \ .$$

Résultant de cette décomposition, les indices de Sobol' élémentaires et globaux sont respectivement définis par :

$$S_I = \frac{V_I}{\mathrm{Var}[Y]} \quad \text{et} \quad \underline{S_I} = \frac{\sum\limits_{J \subset I} V_J}{\mathrm{Var}[Y]} \ .$$

Notre étude porte plus précisément sur les indices d'ordre un et d'ordre deux globaux définis respectivement comme suit :

$$S_i = \frac{V_i}{\mathrm{Var}[Y]} \quad \text{et} \quad \underline{S_{ij}} = \frac{V_i + V_j + V_{ij}}{\mathrm{Var}[Y]} \ .$$

Finalement, notons que $\sum\limits_{I \subset \{1,\dots,d\}, I \neq \emptyset} S_I = 1$, ce qui permet une interprétation directe de la valeur de chaque indice.

Plans répliqués

L'estimation de ces indices repose sur l'utilisation d'un grand nombre d'appels au modèle consistant en un ensemble de points définis par un plan d'expérience. Différents estimateurs sont alors possibles. Dans son papier (1993), Sobol' propose un estimateur "pick-freeze" de ces indices. La procédure d'estimation associée a un coût linéaire, relatif à la dimension de l'espace des paramètres d'entrée, pour estimer tout les indices d'ordre un. Cette dépendance linéaire disparaît en utilisant des plans d'expériences répliqués. Pour ces plans, le nombre d'appels au modèle devient indépendant de la dimension de l'espace des paramètres d'entrée. Une synthèse sur l'utilisation des plans répliqués (appelés "permuted column sampling plans") peut être trouvée dans le papier de Morris *et al.* (2008) où les auteurs utilisent l'approche introduite par McKay (1995). Dans leur papier, Mara *et al.* (2008) combinent des plans répliqués avec des estimateurs "pick-freeze" pour estimer les indices d'ordre un de Sobol'. Cette procédure a été étudiée plus précisemment (propriétés asymptotiques pour les indices d'ordre un) et généralisée dans le papier de Tissot *et al.* (2014) au cas des indices groupés d'ordre deux. Pour ces derniers, la procédure repose sur la réplication de tableaux orthogonaux randomisés. Etant donné que le budget d'appels au modèle des utilisateurs est souvent limité, il est important d'évaluer le nombre minimal d'appels recquis pour obtenir une estimation satisfaisante des indices. Dans le cas où le nombre de points du plan initial n'est pas suffisant pour obtenir des résultats satisfaisants, une procédure itérative peut être utilisée. Le plan initial est itérativement augmenté, les nouvelles réponses du modèle obtenues s'ajoutant aux précédentes. Cette procédure s'applique facilement aux méthodes classiques (Sobol',Saltelli (2002)) contrairement à la méthode répliquée pour laquelle aucune solution n'a encore été proposée dans la littérature.

2 Construction des plans emboîtés

Nous proposons ici de rendre la méthode répliquée itérative, par l'utilisation d'un plan emboîté (appelé aussi plan augmenté), pour estimer les indices de Sobol' d'ordre un où d'ordre deux groupés. La procédure consiste à augmenter le plan emboîté jusqu'à ce que la précision sur les indices estimés satisfasse un critère d'arrêt tel qu'un seuil relatif à la valeur absolue de la différence entre deux estimations consécutives. La méthode de construction du plan emboîté est spécifique à chaque cas. Pour l'estimation des indices d'ordre un, nous pouvons exploiter le plan emboîté introduit par Qian (2009). Ce plan consiste en un hypercube latin avec plusieurs couches. Cet hypercube latin particulier contient une succession d'hypercubes latins emboîtés. En ce qui concerne les indices d'ordre deux, la construction d'un plan emboîté repose sur la construction d'un tableau orthogonal de force deux. La notion de tableau orthogonal a été introduite par Kishen et a été ensuite étendue par Rao (1946) comme suit :

Definition 1 *Un $t-(q,d,\lambda)$ tableau orthogonal ($t \leq d$) est une matrice de taille $\lambda q^t \times d$ dont les éléments apartiennent à un q-sous-ensemble de $\mathbb{N}$ tel que, pour chaque sous-ensemble de t colonnes de la matrice, chaque sous-ensemble à t éléments du q-sous-ensemble apparait dans exactement λ lignes.*

La méthode standard (sans plan emboîté) pour estimer des indices de Sobol' d'ordre deux globaux consiste à construire un tableau orthogonal de force deux avec un grand nombre de niveaux q. La construction d'un $2-(q,d,\lambda)$ tableau orthogonal (tableau orthogonal de force deux) est facilement réalisable lorsque le nombre de niveaux du tableau orthogonal est un nombre premier. Cette construction résulte de la méthode aux différences introduite par Bose et Bush (1952). Malheureusement, étant donné q_1 et q_2 deux nombres premiers distincts, les tableaux orthogonaux de force deux correspondant ne peuvent pas être emboîtés. La méthode proposée par Qian ne peut donc pas être simplement étendue pour construire un tableau orthogonal de force deux emboîté. Des alternatives ont été proposées par Dey (2012) sous la forme de tableaux orthogonaux symétriques et asymétriques. Cependant chacune de ces deux structures possède des contraintes fortes. Dans le cas des tableaux orthogonaux symétriques, les paramètres t, d et q ne peuvent pas être choisis de manière indépendante. Dans le cas des tableaux orthogonaux asymétriques, la discrétisation n'est pas la même pour chaque dimension. Nous proposons ici deux méthodes pour construire un tableau orthogonal de force deux emboîté afin d'estimer les indices de Sobol' d'ordre un et d'ordre deux globaux à l'aide de la procédure itérative décrite précédemment.

- La première méthode est stochastique et repose sur des résultats de théorie des graphes. Une représentation géométrique possible d'un tableau orthogonal de force deux est un ensemble de sous-hypercubes contenu dans l'hypercube unitaire correspondant à l'espace des paramètres d'entrée. L'idée de la méthode est de combler

itérativement les zones lacunaires de l'hypercube. Nous montrons que le problème de la construction d'un tableau orthogonal de force deux emboîté peut se réécrire comme un problème de recherche de clique de taille maximale. La solution de ce problème est due à Bron C. and Kerbosch J.. Dans leur papier, ces auteurs proposent un algorithme pour trouver toutes les cliques d'un graphe. Beaucoup d'améliorations ont été apportées à cet algorithme, la version que nous utilisons est une version parallèlisable développée par D. Matjaz (2013).

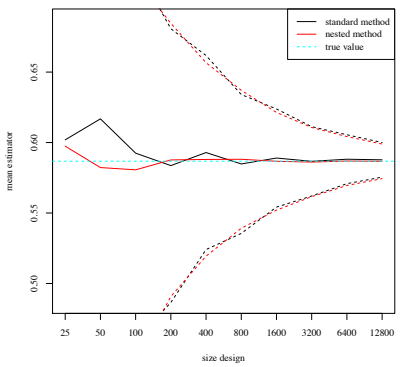
- La deuxième méthode consiste à construire un tableau orthogonal de force deux d'index $\lambda > 1$ puis à rééchantillonner à l'intérieur de chaque cellule par une loi uniforme.

3 Comparaison des méthodes

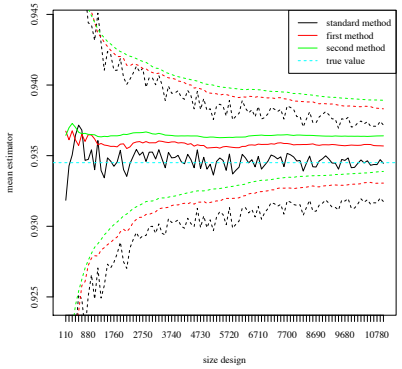
Nous conduisons des tests numériques sur des fonctions classiques afin de comparer les indices d'ordre deux globaux obtenus par chacune des deux méthodes à ceux obtenus par la méthode standard. Pour les indices d'ordre deux et pour chacune des trois méthodes, le nombre de niveaux du tableau est $q = 11$. Sur la figure ci-dessous sont représentées les estimations, en noir, du premier indice d'ordre un et du premier indice d'ordre deux dans le cas de la fonction g-Sobol' définie par :

$$f(X_1, \ldots, X_d) = \prod_{i=1}^{d} \frac{|4X_i - 2| + a_i}{1 + a_i}, \ a_i \geq 0,$$

Nous avons choisi ici $d = 6$ et $(a_1 = 0, a_2 = 0.5, a_3 = 3, a_4 = 9, a_5 = 99, a_6 = 99)$. Nous nous limitons seulement aux premiers indices par souci de taille du document. Ces estimations sont moyennées sur 100 répétitions pour chacun des plans emboîtés. En pointillé sont représentés les intervalles de confiance de niveau 0.95 respectifs à chaque méthode. Nous ajoutons également la valeur théorique de chaque indice.



(a) indice d'ordre un



(b) indice global d'ordre deux

L'approche itérative par plans répliqués est convaincante pour l'estimation des indices d'ordre 1. L'estimation des indices d'ordre 2 présente un biais. Un travail en cours devrait permettre de réduire ce biais. Un autre travail en cours concerne la comparaison des propriétés de recouvrement de l'espace des plans emboîtés obtenus par chacune des deux méthodes à celles du tableau orthogonal de force deux obtenu par la méthode standard. Plusieurs critères sont utilisés tels la mesure de discrepance, le critere d'Eglai, la distance minimale (mindist), la divergence de Kullback- Leibler (Jourdan (2009)) et l'arbre couvrant de poids minimal.

Bibliographie

[1] Rao, C. R. (1946), Hypercubes of strength "d" leading to confounded designs in factorial experiments, Bulletin of the Calcutta Mathematical Society, 38, 67-78.
[2] Bose, R. C. and Bush, K. A. (1952), Orthogonal arrays of strength two and three, The Annals of Mathematical Statistics, 23, 508-524.
[3] Sobol', I. M. (1993), Sensitivity indices for nonlinear mathematical models, Mathematical Modeling and Computational Experiment, 1, 407-414.
[4] McKay, M. D. (1995), Evaluating prediction uncertainty, Los Alamos National Laboratory Report NUREG/CR- 6311, LA-12915-MS.
[5] Saltelli, A. (2002), Making best use of model evaluations to compute sensitivity indices, Computer Physics Communications, 145, 280-297.
[6] Mara, T. A. and Rakoto-Joseph, O. (2008), Comparison of some efficient methods to evaluate the main effect of computer model factors, J. Statist. Comput. Simulation, 78, 167-178.
[7] Morris, M. and Moore, L. M. and McKay, M. D. (2008), Orthogonal Arrays in the Sensitivity Analysis of Computer Models, Technometrics, 50, 205-215.
[8] Jourdan, A. and Franco, J. (2009), A new criterion based on Kullback-Leibler information for space filling designs, https://hal.archives-ouvertes.fr/hal-00375820/.
[9] Qian, P. Z. G. (2009), Nested Latin hypercube designs, Biometrika, 96, 957970.
[10] Dey A. (2012), On the construction of nested orthogonal arrays, Australas. J. Combin., 54, 3748.
[11] Depolli, M. and Konc, J. and Rozman, K. and Trobec, R. and Janežič, D. (2013), Exact Parallel Maximum Clique Algorithm for General and Protein Graphs, J. Chem. Inf. Model., 53, 2217-2228.
[12] Tissot, J. Y. and Prieur, C. (2014), A Randomized Orthogonal Array-based procedure for the estimation of first- and second-order Sobol' indices, To appear in J. Statist. Comput. Simulation.