Proposition de post-doctorat d'un an en 2011

Estimation et utilisation des dérivées pour l'analyse de sensibilité globale

Bertrand Iooss (EDF R&D) - Anne-Laure Popelin (EDF R&D)
Fabrice Gamboa (IMT) - Anestis Antoniadis (UJF)

Localisation : EDF R&D, 6 Quai Watier, 78401 Chatou
Département Management des Risques Industriels
Groupe Fiabilité des Composants et Modélisation des Incertitudes

Financement : Institut de Mathématiques de Toulouse

Contexte du sujet

Ce sujet de post-doctorat est financé par le projet COSTA BRAVA de l'Agence Nationale pour la Recherche. COSTA BRAVA a pour objectif de développer de nouvelles méthodes d'analyse statistique des codes de calcul simulant des dynamiques spatio-temporelles. Le site web du projet http://www.math.univ-toulouse.fr/COSTA_BRAVA contient toutes les informations nécessaires sur ce projet. Dans le cadre de ce projet, EDF R&D coordonne l'axe "Analyse de sensibilité", thématique principale de ce sujet de post-doctorat.

Description du sujet

Les méthodes d'analyse de sensibilité globale sont à présent couramment utilisées dans l'industrie comme outil d'exploration de modèles numériques simulant des phénomènes complexes (Saltelli et al. [13], De Rocquigny et al. [5]). Elles permettent de quantifier l'influence de chaque variable d'entrée (notée $X_i, i = 1, \ldots, d$) sur les différentes sorties d'un modèle (noté $f(\cdot)$). Les méthodes basées sur la variance permettent notamment de ne formuler aucune hypothèse quant à la linéarité/monotonie du modèle vis-à-vis de ses variables d'entrée $\mathbf{X} = (X_1, \ldots, X_d)$ (Sobol [15]). Elles présentent donc des avantages certains par rapport à la méthode traditionnelle, nommée analyse de sensibilité locale, consistant à calculer les dérivées de la sortie du modèle par rapport à chaque variable d'entrée pour un jeu de valeurs nominales des variables d'entrée du modèle (Saltelli et al. [13]). En effet, ce type d'approches suppose la linéarité du modèle dans le domaine considéré et l'absence d'effet sur la sortie du modèle provenant des interactions potentielles entre les variables d'entrée. Les méthodes locales demeurent néanmoins pertinentes, *via* l'approche adjointe (Cacuci [2]), dans les problèmes où le nombre de variables d'entrée d est très grand (typiquement $d > 100$), par exemple dans les modèles météorologiques.

L'analyse de sensibilité globale basée sur la variance permet quant à elle d'obtenir des indices de sensibilité quantitatifs, nommés indices de Sobol, variant entre 0 et 1, et s'interprétant en terme de pourcentage de la contribution d'une variable dans la variance de la sortie du modèle (Saltelli et al. [13]). L'indice de sensibilité du premier ordre S_i donne l'effet de la variable d'entrée X_i, prise seule (*i.e.* sans ses interactions avec d'autres variables), sur la variance de la sortie. Cela revient à estimer la variance d'une espérance conditionnelle $\mathrm{Var}(\mathbb{E}(f(\mathbf{X})|X_i)$. En pratique, on est souvent plus intéressé par l'obtention de l'indice de sensibilité total S_{T_i} de la variable d'entrée X_i (Homma & Saltelli [6]), qui donne l'effet total d'une variable d'entrée sur la sortie (incluant son effet individuel mais aussi les effets de toutes ces interactions avec les autres variables d'entrée).

Pour estimer les indices de Sobol du premier ordre et les indices de Sobol totaux, des méthodes basées sur des échantillons Monte Carlo ont été développées (Saltelli [12]). Malheureusement, pour obtenir des estimations précises des indices de sensibilité, ces méthodes sont extrêmement coûteuses en nombre d'évaluations du modèle (taux de convergence en $\sqrt{N}$ où N est la taille de l'échantillon), d'autant plus que l'algorithme d'estimation des indices de Sobol est particulière-

ment instable. Il n'est pas rare dans les applications que l'estimation d'un indice de Sobol requiert 10000 évaluations de $f(\cdot)$ pour obtenir une précision de 10%. De plus, les évaluations réalisées pour estimer un indice ne sont pas réutilisées pour les autres indices. L'utilisation d'échantillons déterministes de type quasi Monte Carlo (par exemple les séquences LPτ de Sobol) à la place d'échantillons Monte Carlo permet de réduire d'un facteur 10 le coût de ces estimations (Saltelli et al. [14]). Celui-ci demeure néanmoins élevé, d'autant plus dans les problèmes à grande dimension en entrée car le coût en nombre d'évaluations est directement proportionnel à d.

Afin de combler ces lacunes, des travaux récents se sont intéressés au développement de nouveaux estimateurs pour les indices de Sobol du premier ordre. Da Veiga et al. [3] proposent un estimateur asymptotiquement consistant, basé sur des polynômes locaux. Da Veiga & Gamboa [4] proposent quant à eux un estimateur asymptotiquement efficace à partir d'estimations non paramétriques d'intégrales de fonctionnelles de densité. Ces approches semblent prometteuses, mais l'estimation à moindre coût d'indices totaux demeure un axe de recherche ouvert et de première importance. C'est sur ce point spécifique que l'on souhaite se concentrer dans ce sujet de recherche.

Dans un papier récent, Kucherenko et al. [8] ont proposé une première piste de recherche en comparant les indices de Sobol à différentes intégrales des dérivées du modèle $f(\cdot)$. Ils montrent notamment de manière heuristique trois propriétés pouvant s'avérer particulièrement utiles dans la pratique :

$$\mathbb{E}\left|\frac{\partial f}{\partial X_i}\right| \Big/ \sqrt{\mathrm{Var}\left(\frac{\partial f}{\partial X_i}\right)} \propto \frac{S_i}{S_{T_i}} \tag{1}$$

$$S_i \leq \mathbb{E}\left[\left(\frac{\partial f}{\partial X_i}\right)^2\right] \leq S_{T_i} \tag{2}$$

$$\mathbb{E}\left[\left(\frac{\partial f}{\partial X_i}\right)^2\right] \simeq 0.9 S_{T_i} \tag{3}$$

Les mesures utilisant les dérivées sont nommées DGSM ("Derivative-based Global Sensitivity Measures"). Les propriétés (1) et (3) se révèlent correctes pour les fonctions tests de faible dimension effective (Owen [10]), ce qui est souvent le cas avec les modèles numériques. La propriété (2) se révèle correcte pour tout type de fonctions, ce qui permet d'avoir une borne inférieure, potentiellement plus élevée que l'indice de Sobol du premier ordre, pour l'indice de Sobol total. Il est également montré que l'évaluation des indices DGSM par des méthodes de Monte Carlo ou quasi Monte Carlo est bien moins coûteuse que l'évaluation des indices de Sobol. Sobol & Kucherenko [16] ont ensuite obtenu des résultats mathématiques plus précis. Ils ont notamment fourni une borne supérieure pour l'indice de Sobol total. En pratique, l'utilisation des indices DGSM est conditionnée par le calcul des dérivées du modèle, ce qui peut être réalisé par différences finies ou par différentiation automatique (Isukapalli et al. [7]).

Ce sujet de post-doctorat consiste donc à étudier des méthodes de calcul de dérivées plus efficaces pour le calcul des intégrales des indices DGSM. Une première approche serait de réaliser une approximation par decomposition en ondelettes-vaguelettes (Antoniadis & Bigot [1], voir aussi Rao [11] pour le cas des densités). L'idée essentielle est de décomposer la fonction du modèle $f(\cdot)$ dans une base d'ondelettes qui quasi-diagonalise les opérateurs de différentiation. Les différentes intégrales des dérivées du modèle $f(\cdot)$, nécéssaires au calcul des indices DGSM sont alors directement exprimées à l'aide des coefficients de ces décompostions en ondelettes-vaguelettes permettant, on l'espère, un gain en efficacité. Une seconde approche, que nous désirons explorer, serait d'utiliser des méthodes par polynômes locaux. En effet, la plupart des travaux sur la régression non paramétrique par polynômes locaux se concentrent sur l'estimation de la fonction elle-même alors que la méthodologie est adaptée et efficace pour l'estimation des dérivées du modèle sans effort supplémentaire (Liu [9]). Le calcul des intégrales nécessaire à la détermination des indices

DGSM pourrait alors être réalisé en utilisant les techniques issues des travaux de Da Veiga et al. [3], consistant à partitionner l'échantillon des observations en deux parties : l'une pour estimer par polynômes locaux les dérivées de la régression et l'autre pour estimer les intégrales de ces dérivées par des estimateurs empiriques classiques. Bien entendu, ce sujet de recherche est ouvert et bien d'autres pistes pourront être explorées.

Durant son post-doctorat, le chercheur sera amené à tester ses méthodes sur des cas tests à dimension effective variable, puis à traiter des applications réelles issues de besoins d'EDF R&D. Il sera également amené à collaborer avec le Laboratoire Jean Kuntzmann de Grenoble, partenaire d'EDF R&D au sein de COSTA BRAVA, qui travaille sur des applications environnementales lourdes traitées habituellement par approches adjointes.

References

[1] A. Antoniadis and J. Bigot. Poisson inverse problems. *Annals of Statistics*, **34**:5, 2132-2158, 2006.

[2] D.G. Cacuci. *Sensitivity and uncertainty analysis - Theory*. Chapman & Hall/CRC, 2003.

[3] S. Da Veiga, F. Wahl and F. Gamboa. Local polynomial estimation for sensitivity analysis for models with correlated inputs. *Technometrics*, submitted.

[4] S. Da Veiga and F. Gamboa. Efficient estimation of non linear conditional functionals of a density. *Annals of Statistics*, submitted.

[5] E. De Rocquigny, N. Devictor and S. Tarantola, editors. *Uncertainty in industrial practice*. Wiley, 2008.

[6] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of non linear models. *Reliability Engineering and System Safety*, 52:1-17, 1996.

[7] S.S. Isukapalli, A. Roy and P.G. Georgopoulos. Efficient sensitivity/uncertainty analysis using the combined stochastic response surface method and automated differentiation: application to environmental and biological systems. *Risk Analysis*, 20:591-602, 2000.

[8] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides and N. Shah. Monte Carlo evaluation of derivative-based global sensitivity measures. *Reliability Engineering and System Safety*, 94:1135-1148, 2009.

[9] Z-Q. Liu. Multivariate locally weighted polynomial fitting and partial derivative estimation. *Journal of Multivariate Analysis*, 59:187-205, 1996.

[10] A.B. Owen. The dimension distribution and quadrature test functions. *Statistica Sinica*, 13:1-17, 2003.

[11] B.L.S.P. Rao. *Nonparametric estimation of a multivariate probability density by the method of wavelets*. Asymptotics in Statistics and Probability, Festschrift for G.G.Roussas (Ed. M.L.Puri), VSP, The Netherlands, 321-330, 2000.

[12] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communication*,145:280-297, 2002.

[13] A. Saltelli, K. Chan and E.M. Scott. *Sensitivity analysis*. Wiley, 2000.

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

[15] I.M. Sobol. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 55:271-280, 1993.

[16] I.M. Sobol and S. Kucherenko. Derivative based global sensitivity measures and their links with global sensitivity indices. *Mathematics and Computers in Simulation*, 79:3009-3017, 2009.