Planification d'expériences séquentielle pour l'analyse de sensibilité

Loïc Le Gratiet [1,2] & Mathieu Couplet [1] & Bertrand Iooss [1,3] & Luc Pronzato [2]

[1] *EDF R&D, 6 quai Watier, 78 401 Chatou*
[2] *Laboratoire I3S UNS-CNRS, 2000 route des Lucioles, 06903 Sophia-Antipolis*
[3] *Institut de Mathématiques de Toulouse, 31062 Toulouse*

Résumé. Les gros codes de calcul sont communément utilisés en science et en ingénierie pour modéliser des systèmes physiques complexes avec souvent un grand nombre de paramètres d'entrée mal connus. L'analyse de sensibilité globale a pour objectif d'identifier les entrées qui ont le plus d'influence sur la sortie. Les indices de Sobol permettent d'effectuer une telle analyse. Cependant, leur estimation nécessite un grand nombre d'appels au code et ne peut pas être effectuée en un temps raisonnable en pratique. Pour traiter ce problème, un modèle de krigeage est utilisé pour approcher la sortie du simulateur à partir de quelques-unes de ses observations. Ce métamodèle peut ensuite être appelé intensément pour l'estimation des indices. L'utilisation de la variance de krigeage permet de mettre en place des stratégies de planification d'expériences séquentielle efficace pour enrichir le métamodèle. Ces stratégies doivent s'adapter à l'objectif de l'étude (prédiction, optimisation, estimation de probabilité de défaillance,...). Ce travail porte sur la planification séquentielle optimale pour l'estimation des indices de Sobol.

Mots-clés. Analyse de sensibilité, planification séquentielle, krigeage, indices de Sobol.

Abstract. Complex computer codes are widely used in science and engineering to model physical systems. It is common that they have a large number of input parameters. Global sensitivity analysis aims to identify those which have the most important impact on the model output. Sobol indices are a popular tool to perform such analysis. However, their estimations require an important number of simulations and often cannot be processed under reasonable time constraint. To deal with this issue, a kriging model is built to approximate the computer code from few of its observations, and then used to estimate Sobol indices are estimated with it. One of the main strength of kriging models is that they provide through the kriging variance an estimate of the metamodel error. This allows the construction of sequential strategies for improving the metamodel. These strategies depend on the application (prediction, optimization, quantile estimation, estimation of a probability of failure). This work deals with optimal sequential experimental design strategies for the estimation of Sobol indices.

Keywords. Sensitivity analysis, sequential design, kriging, Sobol indices.

1 Introduction

Les codes de calculs ont souvent un nombre important de paramètres d'entrée dont nous voulons mesurer l'influence sur la sortie. Nous nous concentrons ici sur les indices de Sobol qui sont une mesure de sensibilité basée sur une décomposition de la variance. Ces indices sont souvent estimés par des méthodes de Monte-Carlo. Elles permettent de quantifier l'erreur due aux intégrations numériques et elles assurent des propriétés intéressantes comme la normalité asymptotique de l'estimateur. Cependant, elles nécessitent un grand nombre d'appels au code qui est souvent très coûteux en temps de calcul. Pour résoudre ce problème, une approximation du code est souvent faite à l'aide d'un métamodèle. Les indices sont alors estimés à partir de cette approximation.

Dans ce travail, nous utilisons un métamodèle de krigeage. Un de ses avantages est qu'il fourni, au travers de la variance de krigeage, une estimation de l'erreur de métamodèle. Ceci nous permet de quantifier l'erreur d'estimation sur l'indice de Sobol due au méta-modèle. L'utilisation du krigeage pour l'estimation des indices de Sobol (Marrel et al. [1], Oakley et al. [2]) permet de propager l'erreur de métamodèle sur l'erreur d'estimation des indices de Sobol. En revanche, l'erreur numérique due aux intégrations multiples n'est pas prise en compte. Dans (Le Gratiet et al. [6]) il est montré comment prendre en compte ces deux erreurs en combinant un modèle de krigeage avec une méthode d'intégration Monte-Carlo, ce qui permet de les équilibrer.

Supposons maintenant que l'on veuille enrichir le méta-modèle afin de diminuer l'incertitude sur les indices. La question qui nous intéresse porte sur le choix optimal de ces nouveaux points. Il est connu que ces stratégies séquentielles doivent être adaptées à l'estimation des quantités d'intérêts. Ainsi, les nouveaux points choisis seront différents selon que l'on veuille faire de l'optimisation (Jones et al. [4]), de l'estimation de probabilité de défaillance (Bect et al. [5]) ou de la prédiction (Le Gratiet et al. [7]). Jusqu'à présent le cas de l'analyse de sensibilité n'a pas été traité. La principale difficulté est que l'indice de Sobol n'est pas une transformation linéaire du processus Gaussien présent dans le krigeage. Ainsi, on perd la normalité et le calcul de la variance de l'estimateur devient complexe (en particulier, elle dépend des observations du processus). Nous proposons dans ce document deux méthodes pour effectuer une telle planification d'expériences séquentielle.

Le résumé est organisé comme suit. En Section 2 la définition des indices de Sobol est rappelée ainsi qu'un de ses estimateurs Monte-Carlo. En Section 3 nous décrivons les équations du krigeage et présentons l'estimateur des indices couplant krigeage et intégration Monte-Carlo. En Section 4 une première stratégie de planification reposant sur la génération d'instances de processus Gaussien est présentée. Enfin en Section 5, la deuxième stratégie basée sur l'expression analytique de la variance du numérateur de l'estimateur de Sobol est proposée.

Des tests numériques sur les deux stratégies de planification sont en cours et seront présentés lors de la conférence.

2 Les indices de Sobol

Nous présentons ici succinctement la méthode de Sobol pour l'analyse de sensibilité. Considérons l'espace des paramètres d'entrée $Q \in \mathbb{R}^d$. Nous notons $z(X)$, $X \in Q$ la sortie du code de calcul considérée où X est le vecteur des paramètres d'entrée. Pour prendre en compte les incertitudes sur les entrées, X est modélisé comme une variable aléatoire. Soit $X = (X^1, \ldots, X^d)$, l'indice de Sobol du premier ordre du paramètre X^k est défini par:

$$S^k = \mathrm{Var}_X \left(\mathbb{E}_X \left[z(X) | X^k \right] \right) / \mathrm{Var}_X \left(z(X) \right).$$

L'estimation de S^k nécessite donc l'évaluation de différentes intégrales multiples. Afin d'en contrôler l'erreur, nous utilisons une intégration Monte-Carlo présentée ci-dessous. Considérons le couple de variable aléatoire $(X, \tilde{X})$ tel que

$$X = (X^1, \ldots, X^{k-1}, X^k, X^{k+1}, \ldots, X^d),$$

$$\tilde{X} = (\tilde{X}^1, \ldots, \tilde{X}^{k-1}, X^k, \tilde{X}^{k+1}, \ldots, \tilde{X}^d),$$

et $\tilde{X}^i$ indépendant de X^i pour tout $i \neq k$. Nous avons l'égalité $S^k = \mathrm{cov}_X(z(X), z(\tilde{X}))/\mathrm{Var}_X \left(z(X) \right)$. De cette égalité, nous pouvons naturellement déduire l'estimation suivante de S^k à partir d'un échantillon $(X_i, \tilde{X}_i)_{i=1,\ldots,m}$:

$$S^k_m = \frac{\frac{1}{m} \sum_{i=1}^m z(X_i) z(\tilde{X}_i) - \frac{1}{m^2} \sum_{i,j=1}^m z(X_i) z(\tilde{X}_j)}{\frac{1}{m} \sum_{i=1}^m z(X_i)^2 - \left(\frac{1}{m} \sum_{i=1} z(X_i) \right)^2}.$$

Comme présenté précédemment, cet estimateur requiert un grand nombre de particules Monte-Carlo m. C'est pourquoi en pratique nous substituons le code $z(x)$ par un métamodèle. Ici nous utilisons un modèle de krigeage.

3 Indices de Sobol avec modèle de krigeage

Le principe du krigeage est de considérer que notre connaissance *a priori* du code $z(x)$ peut être modélisée par un processur gaussien $Z(x)$ de moyenne $\mathbf{f}'(x)\boldsymbol{\beta}$ et de noyau de covariance $\sigma^2 r(x, \tilde{x})$ où r est un noyau de corrélation. Le noyau de covariance est généralement paramétré par un vecteur $\boldsymbol{\theta}$. Par soucis de clarté nous ne le faisons pas apparaître dans ce document.

Ensuite, nous approchons le code $z(x)$ par un processus gaussien $Z_n(x)$ suivant la distribution prédictive $[Z(x)|Z(\mathbf{D}) = \mathbf{z}^n]$ où $\mathbf{z}^n$ sont les valeurs connues de $z(x)$ sur les points du plan d'expériences $\mathbf{D} = \{x_1^{\mathrm{des}}, \ldots, x_n^{\mathrm{des}}\}$, $x_i^{\mathrm{des}} \in Q$. Nous avons (pour σ^2 connu et une loi *a priori* impropre uniforme pour $\boldsymbol{\beta}$):

$$Z_n(x) \sim \mathrm{PG}_Z \left(m_n(x), s_n^2(x, \tilde{x}) \right),$$

où $m_n(x) = \mathbf{f}'(x)\hat{\boldsymbol{\beta}} + \mathbf{r}'(x)\mathbf{R}^{-1}\left(\mathbf{z}^n - \mathbf{F}\hat{\boldsymbol{\beta}}\right)$,

$$s_n^2(x) = \sigma^2\left(1 - \begin{pmatrix}\mathbf{f}'(x) & \mathbf{r}'(x)\end{pmatrix}\begin{pmatrix}0 & \mathbf{F}' \\ \mathbf{F} & \mathbf{R}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{f}(x) \\ \mathbf{r}(x)\end{pmatrix}\right),$$

$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{z}^n$, $\mathbf{R} = [r(x_i^{\text{des}}, x_j^{\text{des}})]_{i,j=1,\ldots,n}$, $\mathbf{r} = [r(x_i^{\text{des}}, x)]_{i=1,\ldots,n}$ et $\mathbf{F} = [\mathbf{f}'(x_i^{\text{des}})]_{i=1,\ldots,n}$.

Les indices de Sobol peuvent alors être estimés avec :

$$S_{m,n}^k = \frac{\frac{1}{m}\sum_{i=1}^m Z_n(X_i)Z_n(\tilde{X}_i) - \frac{1}{m^2}\sum_{i,j=1}^m Z_n(X_i)Z_n(\tilde{X}_i)}{\frac{1}{m}\sum_{i=1}^m Z_n(X_i)^2 - \left(\frac{1}{m}\sum_{i=1} Z_n(X_i)\right)^2} \tag{1}$$

Notons que cet estimateur comporte à la fois une erreur Monte-Carlo et une erreur de métamodèle. Le contrôle de ces deux erreurs est traité dans l'article (Le Gratiet et al. [7], Le Gratiet [8]). Ici nous nous concentrons sur la diminution de l'incertitude sur $S_{m,n}^k$ – due à l'utilisation du métamodèle – en enrichissant $\mathbf{D}$ d'un nouveau point x_{n+1}^{des}.

4 Planification d'expériences séquentielle pour l'analyse de sensibilité

Soient $(z_{n,l}(x))_{l=1,\ldots,N}$ N réalisations du processus conditionnel $Z_n(x)$. Nous pouvons générer à partir de ces instances un échantillon $(s_{m,n,l}^k)_{l=1,\ldots,N}$ de $S_{m,n}^k$ comme suit :

$$s_{m,n,l}^k = \frac{\frac{1}{m}\sum_{i=1}^m z_{n,l}(X_i)z_{n,l}(\tilde{X}_i) - \frac{1}{m^2}\sum_{i,j=1}^m z_{n,l}(X_i)z_{n,l}(\tilde{X}_j)}{\frac{1}{m}\sum_{i=1}^m z_{n,l}(X_i)^2 - \left(\frac{1}{m}\sum_{i=1} z_{n,l}(X_i)\right)^2}$$

A partir de $(s_{m,n,l}^k)_{l=1,\ldots,N}$ nous pouvons estimer $\text{Var}_Z\left(S_{m,n}^k\right)$ à l'aide d'un estimateur classique de la variance ($\text{Var}_Z()$ désigne la variance selon $Z_n(x)$). De plus, l'estimation de S^k sera obtenue à partir de la moyenne empirique de ce même échantillon.

Supposons maintenant que l'on veuille rajouter un nouveau point x_{n+1}^{des} au design $\mathbf{D}$. L'objectif est d'estimer la diminution de $\text{Var}_Z\left(S_{m,n}^k\right)$ si on conditionne $Z_n(x)$ sur un nouveau point x_{n+1}^{des}. Cette estimation devant s'effectuer sans connaître la valeur de $z(x_{n+1}^{\text{des}})$, nous supposons que $Z_{n+1}(x) \sim \text{PG}_Z\left(m_n(x), s_{n+1}^2(x,\tilde{x})\right)$. A partir de réalisations de $Z_{n+1}(x)$, nous pouvons de nouveau estimer la variance $\text{Var}_Z\left(S_{m,n+1}^k\right)$ de l'estimateur des indices de Sobol (l'indice $n+1$ souligne le fait que l'on rajoute un point au design). La stratégie de planification séquentielle sera alors de choisir parmi un ensemble de points candidats $(x_l^{\text{cand}})_{l=1,\ldots,C}$ celui qui minimise l'estimation de $\text{Var}_Z\left(S_{m,n+1}^k\right)$.

Pour générer des instances de processus gaussiens, nous utilisons la méthode de conditionnement par krigeage couplé à une version propagative de l'échantillonnage de Gibbs (Lantuéjoul et al. [3]). La génération des instance de $Z_{n+1}(x)$ sur $(X_i, \tilde{X}_i)_{i=1,\ldots,m}$ pour un nombre important C de points candidats demeure néanmoins coûteuse (en particulier

parce que m est souvent compris entre 10,000 et 100,000 en pratique). Pour pallier ce problème, nous considérons le processus suivant :

$$Z_{n+1}(x) = r(x, x_{n+1}^{\text{des}}) \left(m_n(x_{n+1}^{\text{des}}) - Z_n(x_{n+1}^{\text{des}}) \right) / r(x_{n+1}^{\text{des}}, x_{n+1}^{\text{des}}) + Z_n(x)$$

qui suit bien la loi prédictive $[Z_n(x)|Z(x_{n+1}^{\text{des}}) = m_n(x_{n+1}^{\text{des}})]$. L'équation donnée précédemment nous permet d'obtenir des instances de $Z_{n+1}(x)$ à partir de celles de $Z_n(x)$ en un temps de calcul négligeable.

5 Une deuxième stratégie de planification d'expériences séquentielle

Nous avons présenté précédemment une approche de planification d'expériences séquentielle permettant de choisir parmi des points candidats celui diminuant le plus fortement l'incertitude sur l'estimation des indices de Sobol. Cette méthode nécessite néanmoins de générer des instances de processus gaussiens sur un grand nombre de points. Ceci étant coûteux numériquement, nous devons restreindre le nombre de points candidats pour la planification séquentielle. Nous sommes également amenés à considérer que $z(x_{n+1}^{\text{des}}) = m_n(x_{n+1}^{\text{des}})$ alors que nous préfèrerions utiliser sa loi complète :

$$z(x_{n+1}^{\text{des}}) \sim \mathcal{N}\left(m_n(x_{n+1}^{\text{des}}), s_n^2(x_{n+1}^{\text{des}}) \right),$$

Conformément à ce qui est proposé dans les méthodes SUR (Bect et al. [5]) où l'objectif de la planification d'expériences est de minimiser l'incertitude sur l'estimation d'une probabilité de défaillance. Le principal point dur qui nous empêche de considérer cette méthode ici est que l'on n'a pas d'estimation analytique de la variance de $S_{m,n}^k$ (1).

En revanche, nous pouvons expliciter la variance du numérateur dans l'expression (1). Dans une seconde approche, nous proposons donc de chercher le point x_{n+1}^{des} apportant la plus forte diminution sur la variance suivante :

$$\text{Var}_Z \left(\frac{1}{m} \sum_{i=1}^{m} Z_n(X_i) Z_n(\tilde{X}_i) - \frac{1}{m^2} \sum_{i,j=1}^{m} Z_n(X_i) Z_n(\tilde{X}_j) \right) = A - B^2$$

où

$$A = \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, \tilde{X}_i, X_j, \tilde{X}_j) + \frac{1}{m^4} \sum_{i,j,k,l=1}^{m} k(X_i, \tilde{X}_j, X_k, \tilde{X}_l) - \frac{2}{m^3} \sum_{i,j,k=1}^{m} k(X_i, \tilde{X}_i, X_j, \tilde{X}_k),$$

$$B = \frac{1}{m} \sum_{i=1}^{m} \left(s_n^2(X_i, \tilde{X}_i) + m_n(X_i) m_n(\tilde{X}_i) \right) - \frac{1}{m^2} \sum_{i,j=1}^{m} \left(s_n^2(X_i, \tilde{X}_j) + m_n(X_i) m_n(\tilde{X}_j) \right),$$

et

$$\begin{aligned} k(x,y,z,t) &= s_n^2(x,y)s_n^2(z,t) + s_n^2(x,z)s_n^2(y,t) + s_n^2(y,z)s_n^2(x,t) \\ &- s_n^2(x,y)m_n(z)m_n(t) - s_n^2(x,z)m_n(y)m_n(t) - s_n^2(x,t)m_n(y)m_n(z) \\ &- s_n^2(y,z)m_n(x)m_n(t) - s_n^2(y,t)m_n(x)m_n(z) - s_n^2(z,t)m_n(x)m_n(y) \\ &+ m_n(x)m_n(y)m_n(z)m_n(t). \end{aligned}$$

Lors de l'ajout d'un nouveau point x_{n+1}^{des} nous avons :

$$s_{n+1}^2(x,\tilde{x}) = s_n^2(x,\tilde{x}) - s_n^2(x,x_{n+1}^{\text{des}})s_n^2(x_{n+1}^{\text{des}},\tilde{x})/s_n^2(x_{n+1}^{\text{des}},x_{n+1}^{\text{des}})$$

et $m_{n+1}(x) = m_n(x) + s_n^2(x,x_{n+1}^{\text{des}})\left(Z_n(x_{n+1}^{\text{des}}) - \begin{pmatrix}\mathbf{F}\\ \mathbf{f}'(x_{n+1}^{\text{des}})\end{pmatrix}\hat{\boldsymbol{\beta}}\right)$.

Ceci nous permet d'évaluer la diminution espérée de l'incertitude si l'on effectue une nouvelle simulation au point x_{n+1}^{des} (notons que cette quantité ne dépend pas de $z_(x_{n+1}^{\text{des}})$.

Remerciements

Le projet a été partiellement financé par le projet ASINCRONE des défis NEEDS (CNRS).

Bibliographie

[1] Marrel A., Iooss B., Laurent B. and Roustant O. (2009), *Calculation of Sobol indices for the Gaussian process metamodel*, Reliability Engineering & System Safety, 79, pp. 229-238.
[2] Oakley J.E. and O'Hagan A. (2004), *Probabilistic sensitivity analysis of complex models a Bayesian approach*, Journal of the Royal Statisitical Society series B, 66, part 3, pp. 751-769.
[3] Lantuéjoul C. and Desassis N. (2012), *Simulation of a Gaussian random vector: A propagative version of the Gibbs sampler*, 9th International Geostatistics Congress, Oslo, Norway, Hune 11. - 15.
[4] Jones D. R., Schonlau M. and Welch W.J. (1998), *Efficient global optimization of expensive black-box functions*, Journal of Global optimization, 13, 4, pp. 455-492.
[5] Bect J., Ginsbourger D., Li L., Picheny V. and Vazquez E. (2012), *Sequential design of computer experiments for the estimation of a probability of failure*, Statistics and Computing, 22(3), pp. 773-793.
[6] Le Gratiet L., Cannamela C. and Iooss B. (2013), *A Bayesian approach for global sensitivity analysis of (multi-fidelity) computer codes*, submitted to SIAM/ASA UQ, arXiv:1307.2223.
[7] Le Gratiet L. and Cannamela C. (2013), *Kriging-based sequential design strategies using fast cross-validation techniques with extensions to multi-fidelity computer codes* , submitted to Technometrics, arXiv:1210.6187.
[8] Le Gratiet L. (2013), *Multi-fidelity Gaussian process regression for computer experiments*, PhD thesis, Université de Paris VII, tel-00866770.