Approche bayésienne pour l'estimation d'indices de Sobol

Benoit Jan [1] & Julien Bect [2] & Emmanuel Vazquez [3] & Pierre Lefranc [4]

[1,2,3] *SUPELEC, Gif-sur-Yvette, France.*
prenom.nom@supelec.fr

[4] *Grenoble Electrical Engineering Lab (G2Elab) – Power Electronics group, St Martin d'Hères, France.*
pierre.lefranc@g2elab.grenoble-inp.fr

Résumé. Le problème considéré est l'estimation des indices de Sobol du premier ordre d'une fonction réelle f coûteuse à évaluer, à partir d'un nombre réduit d'évaluations. Nous nous intéressons à la loi *a posteriori* de ces indices, lorsque f est modélisée par un processus gaussien. Nous montrons qu'il peut être risqué de procéder à une estimation de ces distributions par une approche de type plug-in pour les hyperparamètres du processus gaussien — l'incertitude sur ces hyperparamètres pouvant constituer une part importante de l'incertitude sur les indices de Sobol — et qu'il est préférable d'employer une approche complétement bayésienne. Nos propos sont illustrés sur un exemple académique, puis sur un cas-test issu de l'électronique de puissance.

Mots-clés. processus gaussien, krigeage bayésien, analyse de sensibilité, indices de Sobol

Abstract. We consider the problem of estimating the first-order Sobol indices of a real function f based on few evaluations, when f is expensive to evaluate. We are interested in the posterior distribution of these indices, when f is represented by a Gaussian process. We show that it can be risky to use a plug-in approach with respect to the process hyperparameters for the estimation of these distributions—the uncertainty on these hyperparameters possibly accounting for a major part of the uncertainty on the Sobol indices—and suggest instead to use a fully Bayesian approach. We illustrate the argument with an academic example and a use-case from the power electronics domain.

Keywords. Gaussian process, Bayesian kriging, sensitivity analysis, Sobol indices

1 Introduction

Le problème considéré est l'estimation des indices de Sobol du premier ordre d'une fonction coûteuse à évaluer $f : \mathbb{X} \to \mathbb{R}$, avec $\mathbb{X}$ un pavé de $\mathbb{R}^d$. Ces indices sont définis par

$$S_i(f) = \frac{\mathsf{var}[\mathsf{E}(f(X)|X_i)]}{\mathsf{var}[f(X)]}, \quad i = 1, \ldots, d,$$

où X désigne une variable aléatoire sur $\mathbb{X}$. Dans cet article, nous considérerons que X est uniformément distribuée sur $\mathbb{X}$.

Estimer ces indices par des méthodes de type Monte-Carlo requiert un nombre très important d'évaluations de f (voir par exemple Janon et al. (2012) pour des estimateurs de type Monte-Carlo de ces indices), ce qui n'est pas souhaitable lorsque f est coûteuse à évaluer. Plusieurs auteurs (par exemple Oakley et O'Hagan (2004), Marrel et al. (2009), Storlie et al. (2009)) se sont intéressés à des méthodes permettant d'estimer ces indices de Sobol en faisant intervenir une modélisation de f par un processus gaussien. Dans cette approche, f est vue comme une trajectoire d'un processus aléatoire ξ supposé gaussien conditionnellement à un jeu d'hyperparamètres inconnus $\theta \in \Theta \subset \mathbb{R}^s$. Cette modélisation implique que les indices de Sobol sont des variables aléatoires $S_i(\xi)$, dont on souhaite connaître la loi sachant les valeurs $\xi(x_1), \ldots, \xi(x_n)$ du processus en n points d'observation $x_1, \ldots, x_n \in \mathbb{X}$. Les motivations de nos travaux et nos choix d'implémentation sont exposés dans la section 2. La section 3 présente des résultats numériques obtenus sur un exemple académique, puis sur un cas-test dans le domaine de l'électronique de puissance.

2 Estimation des lois marginales des indices de Sobol

Principe général La modélisation par processus gaussien rend possible la simulation de trajectoires conditionnelles sachant les observations et une valeur de θ : l'idée est alors d'obtenir la distribution des $S_i(\xi)$ par des simulations conditionnelles. L'indice $S_i(\xi)$ peut être approché par un estimateur $\hat{S}_i^m(\xi)$ nécessitant l'évaluation de ξ en $m \gg n$ points de $\mathbb{X}$. La distribution des $S_i(\xi)$ peut alors être approchée par la distribution des $\hat{S}_i^m(\xi)$, que l'on peut obtenir en effectuant des simulations conditionnelles de ξ sur les m points utilisés dans $\hat{S}_i^m(\xi)$. Lorsque la covariance du processus est séparable, il existe des simplifications qui rendent possible la simulation des effets principaux de chacune des variables, comme le proposent Marrel et al. (2009). Dans le cadre de notre article, les covariances utilisées sont non séparables.

L'approche couramment utilisée dans la littérature est une approche de type plug-in, où toutes les opérations sont menées avec une unique valeur $\hat{\theta}$ de θ (obtenue par exemple par maximum de vraisemblance, comme dans Marrel et al. (2009) et Storlie et al. (2009), ou par maximum *a posteriori*, comme dans Oakley et O'Hagan (2004)). L'incertitude sur l'estimation de θ à partir d'un petit nombre d'observations peut constituer une part importante de l'incertitude totale, et utiliser une telle approche comporte donc le risque de donner une fausse idée sur la précision des résultats renvoyés.

Nous proposons une approche complétement bayésienne, qui suppose un *a priori* π_0 sur θ afin d'estimer les lois des $\hat{S}_i^m(\xi)$ intégrées par rapport à la distribution *a posteriori* de θ. Cette approche permet de rendre compte de la méconnaissance des valeurs des hyperparamètres dans la loi *a posteriori* des indices de Sobol.

Détails d'implémentation Nous avons choisi de modéliser la fonction f comme une trajectoire d'un processus gaussien de moyenne constante (inconnue, distribuée uniformément sur $\mathbb{R}$), et de fonction de covariance une covariance de Matérn stationnaire anisotrope, paramétrée par sa variance et ses d paramètres de portée. On pourra se référer à Benassi et al. (2012) pour un choix d'*a priori* π_0 sur les paramètres de la covariance similaire à celui retenu dans le cadre de cet article. L'échantillonnage de la loi *a posteriori* des portées est mené à l'aide d'un algorithme de Metropolis-Hastings indépendant adaptatif inspiré de Ji et Schmidler (2010). La loi *a posteriori* est finalement représentée par J valeurs $\theta_1, \ldots, \theta_J \in \Theta$. Le choix des estimateurs $\hat{S}_i^m$ sera explicité sur les différents exemples de la section 3.

3 Illustrations

Etude d'un cas académique La fonction g de Sobol à d variables est définie sur $[0,1]^d$ par $g(x) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k}$, avec $a_k > 0, \forall k$. Les indices de Sobol du premier ordre de cette fonction peuvent être calculés analytiquement (voir Sobol et al. (2007) et ses références). Sur la figure 1 sont représentées des lois d'indices de Sobol obtenues pour cette fonction en dimension 2, à partir de différents plans d'expériences (deux LHS maximin, de tailles $n = 10$ et $n = 30$). L'estimateur utilisé pour cet exemple calcule les indices de Sobol « discrets » sur une grille régulière de taille 20 par 20.

Pour un nombre d'observations faible ($n = 10$), la loi obtenue avec l'approche plug-in a des modes éloignés des vraies valeurs des indices de Sobol et des queues de distribution petites, tandis que la loi obtenue par l'approche complétement bayésienne possède des queues de distribution plus lourdes, et communique donc des résultats plus prudents.

Lorsque le nombre d'observations est plus grand ($n = 30$), les deux approches donnent des résultats identiques (l'information contenue dans les données observées l'emporte sur la distribution *a priori* des hyperparamètres). Ces différences de comportement entre les deux approches illustrent un risque de l'approche plug-in par rapport à l'approche complétement bayésienne, qui est de sur-estimer la confiance à porter (au résultat de l'étape d'estimation des paramètres et donc) aux indices de Sobol estimés.

Application a un cas-test issu de l'électronique de puissance Le problème posé est celui de l'analyse de sensibilité sur le rendement du convertisseur de puissance représenté schématiquement sur la figure 2. Ce type de convertisseur est utilisé pour des circuits de commande d'interrupteurs de puissance. Le rendement est calculé à l'aide d'un code de simulation prenant en entrée sept facteurs caractérisant le convertisseur (une fréquence de découpage, une capacité, et des distances caractérisant la structure). La simulation du convertisseur fait intervenir un calcul par éléments finis et dure environ deux minutes.

On désigne par $f : [0,1]^7 \to \mathbb{R}$ la sortie du code en fonction des sept facteurs d'entrée. Les observations utilisées sont réparties sur $[0,1]^7$ selon un LHS maximin de taille $n = 10$;

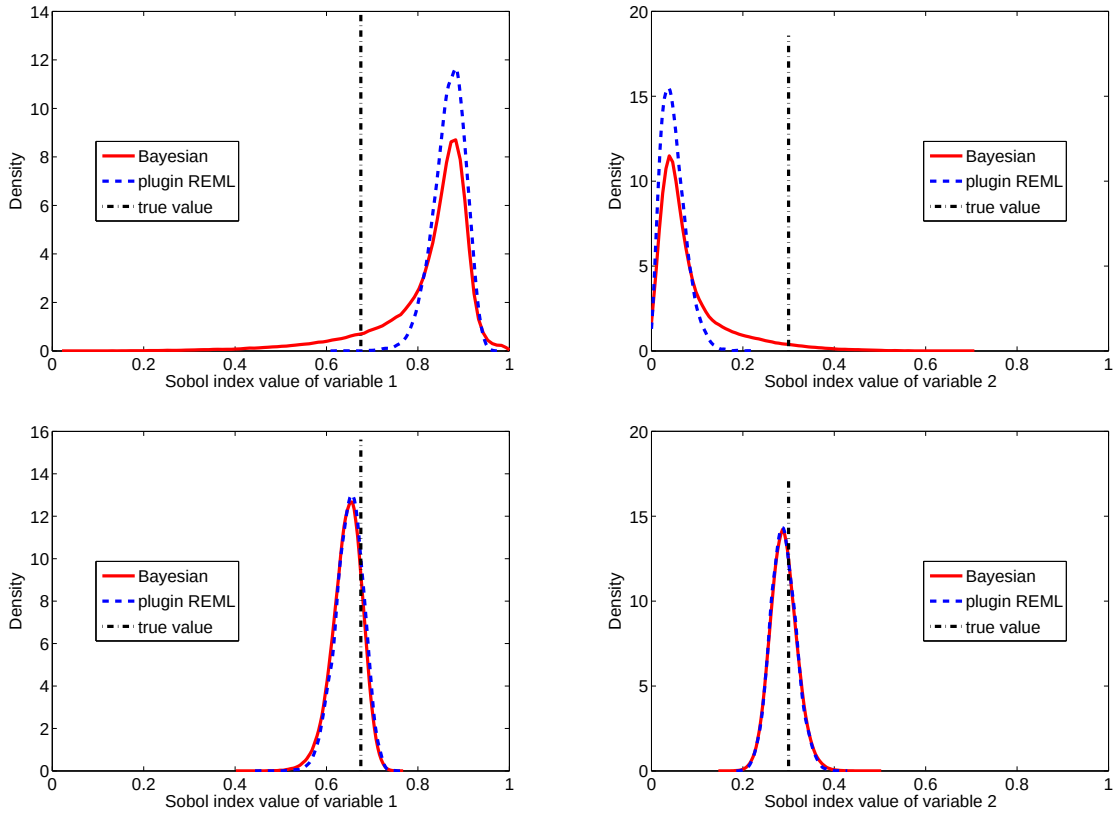

FIGURE 1 – Exemples de distributions d'indices de Sobol sur la fonction g de Sobol de paramètres $a_1 = 1$ et $a_2 = 2$. Les lois représentées sur les figures du haut sont estimées a l'aide de dix observations, celles sur les figures du bas à l'aide de trente observations. Les courbes en trait plein correspondent à la méthode bayésienne, celles en pointillés à la méthode plug-in. Les lignes verticales correspondent aux vraies valeurs.

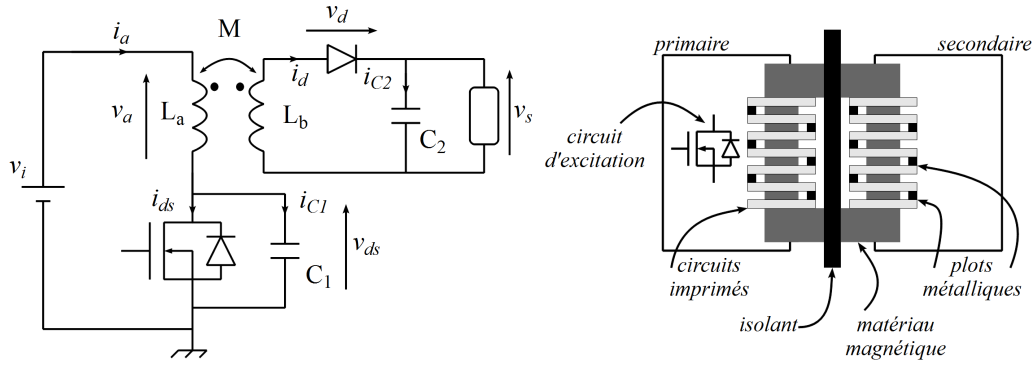

FIGURE 2 – Représentation schématique du convertisseur de puissance. A gauche, une représentation "circuit" du convertisseur, et à droite une description géométrique du transformateur magnétique qu'il contient.

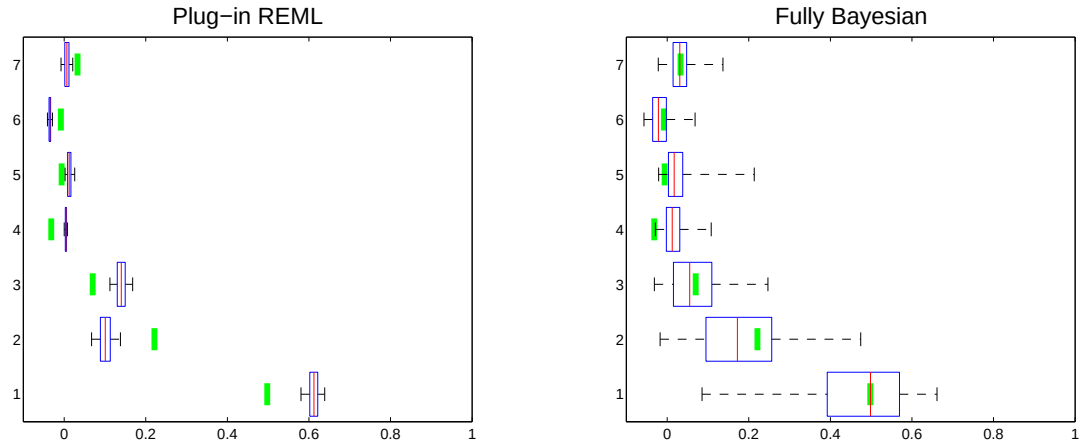

FIGURE 3 – Lois *a posteriori* des indices de Sobol pour le cas-test du convertisseur de puissance, calculées par approche plug-in (à gauche) et par approche complétement bayésienne (à droite) obtenue à partir d'observations réparties sur un LHS maximin de taille dix. Les moustaches contiennent les données allant du quantile à 2,5% au quantile à 97,5%. Les boîtes centrales contiennent les données allant du premier quartile au troisième quartile. Les barres épaisses verticales correspondent aux indices de Sobol $\hat{S}_i^m(f)$ estimés par Monte-Carlo sur la vraie fonction.

la durée de calcul pour obtenir les valeurs de ces observations est d'environ une vingtaine de minutes. L'estimateur $\hat{S}_i^m$ utilisé sur cet exemple est l'estimateur de type Monte-Carlo introduit dans Janon et al. (2012), avec $m = 2000$. Sur la figure 3 sont représentées sous forme de box plots les lois *a posteriori* des indices de Sobol calculées par les méthodes plug-in et complétement bayésienne, à partir du plan d'expériences défini précédemment. Les indices de Sobol de référence sont les valeurs $\hat{S}_i^m(f)$, calculées sur la vraie fonction. Les simulations nécessaires à l'obtention de ces valeurs ont nécessité plus de deux cent cinquante heures de calcul, ce qui est très coûteux et donc non viable en pratique.

Le débordement des lois des indices de Sobol (normalement compris entre 0 et 1) dans les valeurs négatives découle des propriétés de l'estimateur de Monte-Carlo utilisé. Les difficultés que rencontrent les estimateurs de Monte-Carlo classiques pour estimer des indices de Sobol proches de zéro sont évoquées dans Owen (2012). Dans le cas plug-in, sur cet exemple, l'intervalle de crédibilité à 95% caractérisant la loi *a posteriori* de l'indice de Sobol ne contient jamais la valeur de référence. Dans le cas complétement bayésien, les boîtes centrales contiennent la valeur de référence pour plus de la moitié des variables, et dans les autres cas les intervalles de crédibilité à 95% des lois *a posteriori* contiennent les valeurs de référence ou en sont proches (pour la variable 4) . Sur cet exemple encore, la loi *a posteriori* obtenue par l'approche bayésienne est donc bien plus prudente que celle obtenue par l'approche plug-in, qui est très optimiste malgré une importante erreur d'estimation.

4 Conclusions et perspectives

Sur les deux exemples considérés, l'approche de type plug-in a tendance à sous-estimer l'erreur d'approximation faite lors de l'estimation des indices de Sobol, alors que l'approche bayésienne fournit des lois beaucoup plus prudentes (dont les queues de distribution sont plus lourdes). Les véritables taux de couverture des différents intervalles de crédibilité donnés par les deux approches n'ont pas été abordés ici et devront faire l'objet de simulations pour juger de la pertinence des différents intervalles proposés. Le choix de l'*a priori* π_0 pourra être amené à évoluer dans des travaux futurs, et nous envisageons l'implémentation de méthodes séquentielles pour traiter le cas de fonctions dont les espaces d'entrée seront de dimension élevée par rapport à celles étudiées dans cet article (de l'ordre de plusieurs dizaines).

Bibliographie

[1] Benassi, R., Bect, J. et Vazquez, E. (2012), Optimisation bayésienne par méthodes SMC, 44èmes journées de Statistique, Bruxelles, 21–25 mai 2012.

[2] Janon, A., Klein, T., Lagnoux, A., Nodet, M. et Prieur, C. (2012), Asymptotic normality and efficiency of two Sobol index estimators, Preprint, `http://hal.inria.fr/hal-00665048`.

[3] Ji, C. et Schmidler, S. C. (2010), Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection, *Journal of Computational and Graphical Statistics*, à paraître.

[4] Marrel, A., Iooss, B., Laurent, B. et Roustant, O. (2009), Calculations of Sobol indices for the Gaussian process metamodel, *Reliability Engineering and System Safety*, 94(3), 742–751.

[5] Oakley, J. E. et O'Hagan, A. (2004), Probabilistic sensitivity analysis of complex models : a Bayesian approach, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 66(3), 751–769.

[6] Owen, A. B. (2012), Better estimation of small Sobol' sensitivity indices, Preprint, `http://arxiv.org/abs/1204.4763`.

[7] Sobol, I. M., Tarantola, S., Gatelli, D., Kucherenko, S. S. et Mauntz, W. (2007), Estimating the approximation error when fixing unessential factors in global sensitivity analysis, *Reliability Engineering and System Safety*, 92(7), 957-960.

[8] Storlie, C. B., Swiler, L. P., Helton, J. C. et Sallaberry, C. J. (2009), Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models, *Reliability Engineering and System Safety*, 94(11), 1735–1763.