Application de la régression non paramétrique multivariée à l'analyse de sensibilité

Lamia Djerroud [1,a] & Tristan Senga Kiessé [2,b] & Smail Adjabi [1,c]

[1] *Unité de recherche LaMOS, Université de Bájaia, Béjaia, Algerie*
[2]*INRA, UMR 1069 Sol Agro et hydrosystème Spatialisation, F-35000 Rennes, France*
[a] *djerroudlamia@live.fr;* [b] *tristan.sengakiesse@rennes.inra.fr;* [c] *adjabi@hotmail.com*

Résumé. Ce travail est une contribution à l'estimation des indices de sensibilité utilisés pour évaluer l'influence de la variance des variables d'entrée sur la variance de la sortie du modèle de régression. Un estimateur à noyau mixte multivarié est étudié car, jusqu'à présent, les entrées discrètes et continues ont été considérées séparément dans l'estimation des indices de sensibilité par la méthode à noyau. En utilisant la fonction test Ishigami, nous comparons les expressions analytiques de certains indices de sensibilité Sobol calculés pour 3 types de paramètres d'entrée: mixtes, discrets et continus.

Mots-clés. Analyse de variance, Noyau associé, Régression non paramétrique, Analyse de sensibilité, Indice de Sobol.

Abstract. This paper is interested in estimating sensitivity indices useful to evaluate the contribution of the inputs variation to the variance of the regression model. A multivariate mixed kernel estimator is investigated since, until now, discrete and continuous inputs have been considered separately in kernel estimation for sensitivity indices. By using the Ishigami test function, analytical expressions of some Sobol sensitivity indices are expressed for mixed inputs, in comparison to discrete and continuous cases.

Keywords. Analysis of variance, Associated kernel, Nonparametric regression, Sensitivity analysis, Sobol indice.

1 Introduction

Considérons un modèle de régression non-paramétrique multivariée de la forme $Y = f(X_1, X_2, \ldots, X_d)$ avec $Y \in \mathbb{R}$ la variable de sortie et $X_j \in \mathbb{T}$, $j = 1, ..., d$ les variables d'entrée. Les méthodes d'analyse de sensibilité d'un modéle ont pour but de déterminer quelles sont les variables d'entrée qui influencent le plus la variable de sortie afin, entre autres, de réduire le modèle (Sobol (2001)). Les indices de sensibilité sont déterminés

en se basant sur la décomposition par analyse de variance de $f(X_1, X_2, \ldots, X_d)$ en somme de fonctions élémentaires

$$Y = f_0 + \sum_{i=1}^{k} f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \ldots + f_{12\ldots k}(X_1, X_2, \ldots, X_k), \tag{1}$$

avec $f_0 = \mathbb{E}(Y)$, $f_i = \mathbb{E}(Y|X_i) - f_0$, $f_{ij} = \mathbb{E}(Y|X_i, X_j) - f_i - f_j - f_0, \ldots$ La décomposition (1) permet d'obtenir la décomposition suivante de la variance (Sobol (2001))

$$\mathbb{V}(Y) = \sum_{i=1}^{k} \mathbb{V}_i + \sum_{i<j} \mathbb{V}_{ij} + \ldots + \mathbb{V}_{12\ldots k}, \tag{2}$$

avec

$$\mathbb{V}_i = \mathrm{Var}\{\mathbb{E}(Y|X_i)\}, \quad \mathbb{V}_{ij} = \mathrm{Var}\{\mathbb{E}(Y|X_i, X_j)\} - \mathbb{V}_i - \mathbb{V}_j, \ldots$$

À partir de (2) les indices de sensibilité s'expriment comme suit :

$$S_i = \frac{\mathbb{V}_i}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{\mathbb{V}_{ij}}{\mathrm{Var}(Y)}, \quad \cdots .$$

Ces indices deviennent difficiles à interpréter lorsque le nombre de variables d'entrée d augmente d'où l'introduction de l'indice de sensibilité total de Homma et Saltelli (1996) qui exprime l'effet total d'une variable d'entrée sur la sortie du modèle

$$ST_i = S_i + \sum_{j \neq i} S_{ij} + \sum_{j \neq i, k \neq i, j<k} S_{ijk} + \cdots = 1 - \frac{\mathrm{Var}\{\mathbb{E}(Y|X_{-i})\}}{\mathrm{Var}(Y)},$$

où $\mathrm{Var}\{\mathbb{E}(Y|X_{-i})\}$ est la variance de l'espérance de Y conditionnellement à toutes les variables sauf X_i.

Dans ce travail nous nous intéressons à l'évaluation des indices de sensibilité pour des données d'entrée mixtes, i.e. continues et discrètes, en utilisant un estimateur à noyau du type Nadaraya (1964) et Watson (1964) pour la régression (cf. Zhang, King et Shang (2013) pour la régression sur des données mixtes). Ainsi, nous poursuivons les travaux de Luo, Lu et Xu (2014) sur un estimateur à noyau continu, i.e. $\mathbb{T} = \mathbb{R}$, de la décomposition en (1) et ceux de Senga Kiessé et Ventura (2016) sur un estimateur à noyau discret, i.e. $\mathbb{T} = \mathbb{N}$, de cette même décomposition. Les expressions analytiques d'indices de sensibilité de Sobol de plusieurs ordres sont fournies pour des variables d'entrée mixtes en utilisant la fonction test d'Ishigami, en comparaison avec les cas de données d'entrée continues et discrètes. De plus, deux choix de fenêtres sont appliquées pour l'estimation à noyau : la validation croisée et l'approche bayésienne.

2 L'estimateur à noyau de le fonction de régression multivariée

Considérons $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ une séquence indépendante et identiquement distribuée de vecteurs aléatoires définis sur $\mathbb{T}^d \times \mathbb{R}$ sachant que $m(\cdot) = \mathbb{E}(Y^k|\mathbf{X}^k = \cdot)$. Soit $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{T}^d$ le vecteur cible et $\mathbf{H} = \mathbf{Diag}(h_{11}, \ldots, h_{dd})$ la matrice de lissage avec $h_{jj} > 0$ tel que $\mathbf{H} \equiv \mathbf{H}_n$ tend vers la matrice nulle $\mathbf{0_d}$ quand $n \to \infty$. L'estimateur multivarié de la fonction de régression $\widehat{m}_n^d$ de m en utilisant le noyau discret $K_{x_j,h_{jj}}$ avec $(x_j, h_{jj}) \in \mathbb{T}(\subseteq \mathbb{Z}) \times (0, \infty)$ est donné par :

$$\widehat{m}_n^d(\mathbf{x}, \mathbf{H}) = \sum_{i=1}^{n} \frac{y_i K_{\mathbf{x},\mathbf{H}}(\mathbf{x}^i)}{\sum_{l=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{x}^l)},$$

où le noyau associé multivarié $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est le produit des noyaux associés univariés $K^{[j]}_{x_j,h_{jj}}$ lié à une variable aléatoire discrète $\mathcal{K}^{[j]}_{x_j,h_{jj}}$ sur $\mathbb{S}_{x_j,h_{jj}}$ tel que

$$x_j \in \mathbb{S}_{x_j,h_{jj}} \ \ (A1), \quad \lim_{h_{jj}\to 0} \mathbb{E}(\mathcal{K}^{[j]}_{x_j,h_{jj}}) = x_j \ \ (A2), \quad \lim_{h_{jj}\to 0} \mathrm{Var}(\mathcal{K}^{[j]}_{x_j,h_{jj}}) = 0 \ \ (A3).$$

Le noyau associé multivarié $K_{\mathbf{x},\mathbf{H}}$ lié à la variable aléatoire discrète $\mathcal{K}_{\mathbf{x},\mathbf{H}}$ de support $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{j=1}^{d} \mathbb{S}_{x_j,h_{jj}}$ est une fonction de masse de probabilité satisfaisant:

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \ \ \mathbb{E}(\mathcal{K}_{\mathbf{x},\mathbf{H}}) = \mathbf{x} + \mathbf{U}(\mathbf{x}, \mathbf{H}), \ \ \mathbb{C}\mathrm{ov}(\mathcal{K}_{\mathbf{x},\mathbf{H}}) = \mathbf{B}(\mathbf{x}, \mathbf{H}),$$

où $\mathbf{U}(\mathbf{x}, \mathbf{H}) = (u_1(\mathbf{x}, \mathbf{H}), \ldots, u_d(\mathbf{x}, \mathbf{H}))^\top$ et $\mathbf{B}(\mathbf{x}, \mathbf{H}) = (b_{ij}(\mathbf{x}, \mathbf{H}))_{i,j=1,\ldots,d}$ convergent, respectivement, vers le vecteur nul $\mathbf{0}$ et la matrice nulle $\mathbf{0_d}$ quand $\mathbf{H} \to \mathbf{0_d}$ (Sobom et Kokonendji (2016)).

2.1 L'estimateur à noyau de la décomposition (ANOVA)

Pour déterminer l'estimateur de la décomposition donné par la formule (2) nous estimons les fonctions élémentaires de la décomposition de $f(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, ..., x_d)$ par la méthode du noyau multivarié.

D'abord, l'estimateur de f_0 est donné par $\hat{f_0} = (1/n)\sum_{l=1}^{n} y_l$. Puis, l'estimateur de f_i peut s'écrire comme suit

$$\hat{f_i}(x_i, h_i) = \frac{1}{n}\sum_{l=1}^{n} K_{x_i,h_i}(x_{il}) y_l - \frac{1}{n}\sum_{l=1}^{n} y_l = \frac{1}{n}\sum_{l=1}^{n} \mathbb{K}_{x_i,h_i}(x_{il}) y_l,$$

avec $\mathbb{K}_{x_i,h_i} = K_{x_i,h_i}(x_{il}) - 1$. De la même manière on obtient l'estimateur de f_{ij}:

$$\hat{f_{ij}}(x_i, x_j, h_i, h_j) = \frac{1}{n}\sum_{l=1}^{n} \mathbb{K}_{x_i,x_j,h_i,h_j}(x_{il}, x_{jl}) y_l,$$

où $\mathbb{K}_{x_i,x_j;h_i,h_j} = K_{x_i,x_j;h_i,h_j}(x_{il}, x_{jl}) - K_{x_i,h_i}(x_{il}) - K_{x_j,h_j}(x_{jl}) - 1$. Et ainsi de suite pour les autres termes $\widehat{f_{\mathbf{I}}}$. Ensuite, sous les hypothèses (A2)-(A3), nous sommes en mesure d'établir quelques propriétes asymptotiques de l'estimateur $\hat{f_i}$ de f_i. Nous montrons que :

$$\text{MSE}(x_i) = \text{Bias}^2\{\widehat{f_i}(x_i; h_{ii})\} + \text{Var}\{\widehat{f_i}(x_i; h_{ii})\} \to 0 \text{ quand } h_{ii} \to 0 \text{ et } n \to \infty,$$

où MSE désigne l'erreur quadratique moyenne. Puis, nous établissons ce résultat :

Proposition 1 *Pour tout $x_i \in \mathbb{T}$ et $h_{ii} > 0$, l'estimateur à noyau $\widehat{f_i}$ satisfait:*

$$\widehat{f_i}(x_i; h_{ii}) \xrightarrow{p.s.} f_i(x_i) \text{ quand } n \to \infty \text{ et } h_{ii} \to 0.$$

Par la suite, l'estimation des termes de la variance se présente comme suit

$$\widehat{\mathbb{V}}(Y) = \frac{1}{n}\sum_{l=1}^{n} y_l^2 - \widehat{f_0}^2, \ \widehat{\mathbb{V}}_i = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_i}(x_i; h_{ii})\}^2, \ \widehat{\mathbb{V}_{ij}} = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_{ij}}(x_i, x_j; h_{ii}, h_{jj})\}^2, \ldots$$

Finalement, d'après la décomposition de la variance totale de Y donnée par la formule (2) les estimateurs des indices de sensibilité sont

$$\widehat{S_i} = \frac{\widehat{\mathbb{V}}_i}{\widehat{\mathbb{V}}(Y)}, \ S_{ij} = \frac{\widehat{\mathbb{V}_{ij}}}{\widehat{\mathbb{V}}(Y)}, \ \cdots, \widehat{ST_i} = 1 - \frac{\widehat{\mathbb{V}_{-i}}}{\widehat{\mathbb{V}}(Y)}.$$

3 Analyse de la fonction de test Ishigami

Dans cette section, nous évaluons la performance de la méthode de régression non-paramétrique à noyau associé mixte pour l'estimation des indices de Sobol pour la fonction d'Ishigami donnée par:

$$y = m(x_1, x_2, x_3) = sin(x_1) + 5\, sin^2(x_2) + 0.1\, x_3^4 sin(x_1),$$

où $x_i, i = 1, 2, 3$, sont des paramètres d'entrée répartis uniformément sur $\mathbb{T}$ (Ishigami et Homma (1990)). Dans cette étude, le cas mixte considère le paramètre $x_1 \in \mathbb{T} = \{-3, -2, -1, 0, 1, 2, 3\}$ et $x_2, x_3 \in \mathbb{T} = [-\pi, \pi]$.

Deux noyaux ont été considérés: le noyau gaussien et le noyau discret triangulaire symétrique donné par:

$$T_{a;x,h}(y) = \frac{(a+1)^h - |y-x|^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \forall y \in \mathbb{S}_x = \{x, x \pm 1, ..., x \pm a\}, \ a \in \mathbb{N},$$

tel que $x \in \mathbb{T}$ est la cible et $h > 0$ le paramètre de lissage (voir Kokonendji, Senga Kiessé et Zocchi (2007)); notons qu'en pratique, on utilise $a = 1$. Pour le choix de la

Indices de Sobol	Cas continu	Cas discret	Cas mixte
S_1	0.40	0.42	0.39
S_2	0.29	0.19	0.31
S_{13}	0.31	0.39	0.30

Table 1: Valeurs analytiques des indices de sensibilité dans les cas continu, discret et mixte pour la fonction d'Ishigami, $S_3 = S_{12} = S_{23} = S_{123} = 0$.

	n	$\overline{\hat{S}}_1$	$\overline{\hat{S}}_2$	$\overline{\hat{S}}_3$	$\overline{\hat{S}}_{12}$	$\overline{\hat{S}}_{13}$	$\overline{\hat{S}}_{23}$	$\overline{\hat{S}}_{123}$
Validation croisée	250	0.352	0.265	0.038	0.035	0.253	0.229	-0.200
	500	0.350	0.258	0.017	0.018	0.240	0.131	-0.066
	1000	0.351	0.247	0.010	0.007	0.237	0.063	0.015
Approche bayésienne	250	0.363	0.230	0.016	0.026	0.236	0.078	0.006
	500	0.423	0.206	0.002	0.014	0.269	0.072	0.015
	1000	0.394	0.235	0.009	0.012	0.238	0.051	0.006

Table 2: Estimations des indices de sensibilité des paramètres d'entrée mixtes pour la fonction d'Ishigami

matrice des fenêtres optimales, deux méthodes ont éte considérés: la validation croisée et une approche bayesienne. Les résultats de simulation sont obtenus par la méthode de Monte Carlo pour $N_{sim} = 100$ répétitions et différentes tailles d'échantillions n.

Table 1 présente les valeurs des indices de sensibilité calculées analytiquement dans les cas continu, discret et mixte pour la fonction d'Ishigami, tandis que Table 2 présente les moyennes des estimations des indices de sensibilité d'ordre 1 données par :

$$\overline{\hat{S}_i} = \sum_{l=1}^{N} (1/N)\hat{S}_i^{(l)}, \quad i = 1, 2, 3.$$

La performance des estimateurs est évaluée par la moyenne de l'erreur absolue

$$\overline{MAE}(S_i) = (1/N_{sim}) \sum_{l=1}^{N_{sim}} |S_i^{(l)} - \hat{S}_i|.$$

Premièrement, on vérifie que l'ordre d'influence (S_i) des paramètres dans le cas mixte est semblable au cas discret et continu (Table 1). Ensuite, il apparait que le choix de fenêtres par l'approche bayesienne produit de meilleurs résultats que par la validation croisée pour l'estimation des indices de sensibilité sauf pour S_2 (Tables 2 et 3). Cependant le temps d'exécution de l'approche bayésienne est plus long. Finalement, l'on retiendra

	n	$\overline{MAE}_1$	$\overline{MAE}_2$	$\overline{MAE}_3$	$\overline{MAE}_{12}$	$\overline{MAE}_{13}$	$\overline{MAE}_{23}$	$\overline{MAE}_{123}$
Validation croisée	250	**0.040**	**0.049**	0.037	0.037	**0.049**	0.224	0.203
	500	0.035	**0.062**	0.018	0.024	**0.054**	0.128	0.073
	1000	0.038	**0.069**	0.008	0.013	0.058	0.068	0.018
Approche bayésienne	250	0.070	0.112	**0.016**	**0.029**	0.063	**0.078**	**0.036**
	500	**0.031**	0.106	**0.008**	**0.014**	0.058	**0.072**	**0.015**
	1000	**0.007**	0.077	**0.002**	**0.012**	**0.030**	**0.051**	**0.006**

Table 3: Valeur moyenne du critère MAE pour l'estimation des indices de sensibilité d'entrée mixtes pour la fonction d'Ishigami

aussi que le choix d'un estimateur adapté au type de données (discrètes, continues ou mixtes) influence la précision de l'estimation obtenue.

Bibliographie

[1] Homma T., et Saltelli A.(1996). Importance measures in global sensitivity analysis of nonlinear models, *Reliability Engineering & System Safety*, 52(1), 1-17.

[2] Ishigami,T. et Homma, T. (1990). An importance qualification technique in uncertainty analysis for computer models. *In:Proceedings of the ISUMA90, first international symposium on uncertainty modelling and analysis, University of Maryland*, p.398–403.

[3] Kokonendji, C. C., Senga Kiessé, T. et Zocchi, S S.(2007), Discrete triangular distributions and non-parametric estimation for probability mass function, *Journal of Nonparametric Statistics*, 8(6), 241-254.

[4] Nadaraya, E.A. (1964), On estimating regression, *Theory of Probability and its Applications*, 9, 141-142.

[5] Luo, X., Lu, Z. et Xu, X.(2014), Non-parametric kernel estimation for the ANOVA decomposition and sensitivity analysis.*Reliability Engineering and System Safety*, 130, 140–148.

[6] Senga Kiessé, T. et Ventura, A. (2016), Discrete non-parametric kernel estimation for global sensitivity analysis, *Reliability Engineering & System Safety*, 46, 47-54.

[7] Sobom M. S. et Kokonendji, C. C. (2016). Effects of associated kernels in nonparametric multiple regressions, arXiv preprint arXiv:1502.01488.

[8] Sobol, I. M., (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*. 55(1), 271-280.

[9] Watson, G. S. (1964), Smooth regression analysis, *Sankhya Ser. A*, 26, 359-372.

[10] Zhang, X., King, M. L. et Shang, H. L. (2013), Bayesian bandwidth selection for a nonparametric regession model with mixed types of regressors(No.13/13). *Monash University, Department of Econometrics and Business Statistics*.