Discrete and continuous nonparametric kernel estimations for global sensitivity analysis

Tristan Senga Kiessé [1,a] & Andy Andrianandraina [1,b]

[1] *L'Université Nantes Angers Le Mans (LUNAM)*
Chaire Génie Civil Eco-construction
Institut de Recherche en Génie Civil et Mécanique GeM UMR-CNRS 6183
58 rue Michel Ange, 44600 Saint-Nazaire, France
[a]*tristan.sengakiesse@univ-nantes.fr* ; [b]*andi.andrianandraina@univ-nantes.fr*

Résumé. Ce travail porte sur la méthode d'estimation à noyau discret dans le cadre de l'analyse de sensibilité d'un modèle f visant à évaluer l'influence des variables d'entrée discrètes X sur la variable réponse Y. En effet, l'estimation à noyau discret est maintenant connue pour être adaptée au lissage des distributions de données à support discret. Cependant, dans le cadre de l'analyse de sensibilité, seule l'estimation à noyau continu a été étudiée jusqu'à récemment pour évaluer l'influence de variables d'entrée continues comme discrètes. Ici, l'approche à noyau discret est utilisée pour construire un estimateur non-paramétrique du modèle $Y = f(X)$ décomposé par analyse de variance. Des simulations sur la fonction test d'Ishigami et sur un cas d'étude issu du domaine de l'agriculture montrent l'intérêt de l'approche par noyau discret en comparaison avec l'approche par noyau continu à travers l'estimation des indices de sensibilité de Sobol. Pour des paramètres d'entrée discrets qui sont moyennement ou très influents, l'approche discrète estime mieux la contribution de leur variance sur la variance totale du modèle par rapport à l'approche continue.

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

Abstract. Nowadays the discrete kernel estimation is known to be suitable for smoothing discrete functions. This work investigates the discrete kernel approach as a metamodeling approach within the framework of sensitivity analysis (SA) for evaluating the influence of discrete input variables X on a model output Y. Indeed, until now only the continuous kernel approach is applied for both discrete and continuous input variables in SA frame. The discrete kernel approach is used for building a nonparametric estimator of ANOVA decomposition of the model $Y = f(X)$. Some simulations on a test function analysis and on a real case study from agricultural revealed the discrete kernel estimation to be competing to continuous kernel estimation for estimating the Sobol sensitivity indices of discrete input variables. The discrete kernel approach outperforms the continuous kernel one for evaluating the contribution of the moderate or most influential discrete parameters.

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

1 Introduction

The sensitivity analysis method aims at quantifying the influence of an input variable X_i on the output Y under a given model $Y = f(X_1, X_2, \ldots, X_d)$. Based on analysis of variance (ANOVA) decomposition, sensitivity indices useful for evaluating the effects of each input X_i are calculated. Thus, the contribution of input variables to the variance of Y is measured through sensitivity indices given by Sobol (2001) as follows:

$$S_i = \frac{\mathrm{Var}\{\mathbb{E}(Y|X_i)\}}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{\mathrm{Var}\{\mathbb{E}(Y|X_i, X_j)\}}{\mathrm{Var}(Y)}, \ldots \tag{1}$$

Then, the total effect of X_i is measured by

$$ST_i = S_i + \sum_{j \neq i} S_j + \cdots + S_{12\ldots d}.$$

From (1) the direct estimation of conditional expectation $\mathbb{E}(Y|X_i)$ provides an estimate of the main effect sensitivity measure S_i, where $\mathbb{E}(Y|X_i)$ is classically known to be the best approximation of the function $f(\cdot)$ in regression context. Various statistical tools as splines, generalized linear or additive model, polynomial are useful in a metamodeling approach for providing an adjustment of a model (Iooss, 2011). From the point of view of nonparametric smoothing, some methods as the continuous kernel-based estimation (Rosenblatt, 1969) or more recently the State-Dependent Parameter estimation (Ratto et al., 2007) are good choices for estimating $\mathbb{E}(Y|X_i)$. About these two estimation methods, Luo et al. (2014) have recently shown that continuous kernel estimation is equal or better than the SDP estimation in term of performance. However, until now in literature the continuous kernel estimation is evenly applied on continuous input variables as on discrete ones.

Nowadays the discrete kernel estimation proposed by Kokonendji and Senga Kiessé (2011) is known to be suitable than the use of continuous kernels for smoothing discrete functions such as count regression functions on a discrete support $\mathbb{T}$ such as $\mathbb{N}$, the set of positive integers, or $\mathbb{Z}$, the set of integers. This work investigates the discrete kernel estimation for metamodeling within the context of global SA where the input variables X_i are discrete.

2 Nonparametric discrete triangular regression

Assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are n independent copies of (X, Y) defined on $\mathbb{T}(\subseteq \mathbb{Z}) \times \mathbb{R}$. We are interested in the nonparametric regression model

$$Y = m(X) + \epsilon,$$

where $m(\cdot) = \mathrm{E}(Y|X = \cdot)$ is an unknown regression function and the random covariate X is independent of the unobservable error variable ϵ's assumed to have zero mean and finite variance. From the well known works of Nadaraya (1964) and Watson (1964) in continuous kernel estimation, Abdous et al. (2012) proposed a discrete nonparametric estimator of m which is defined as follows. For a fixed point $x \in \mathbb{T}$ and a smoothing parameter $h > 0$, the discrete nonparametric estimator of m is defined by:

$$\widehat{m}_n(x;h) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)}, \tag{2}$$

where the arbitrary sequence of smoothing parameters $h = h(n) > 0$ fulfills $\lim_{n\to\infty} h(n) = 0$ and the *discrete associated kernel* $K_{x,h}(\cdot)$ is a probability mass function (pmf) of random variable (rv) $\mathcal{K}_{x,h}$ with support $\mathbb{S}_x$ such as:

$$x \in \mathbb{S}_x \ \ (A1), \quad \lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x \ \ (A2), \quad \lim_{h\to 0} \mathbb{V}(\mathcal{K}_{x,h}) = 0 \ \ (A3).$$

These three assumptions are fulfilled by both continuous and discrete kernels (Senga Kiessé et al., 2014). One of the main issues of the discrete kernel method is the optimal bandwidth choice and the discrete kernel choice are both equally important.

Concerning discrete kernel choice, we present the discrete symmetric triangular kernel satisfying assumptions (A1) to (A3) such that the corresponding estimator $\widehat{m}_n$ has good asymptotic properties.

Symmetric discrete triangular kernel. For $(a, x) \in \mathbb{N} \times \mathbb{T}$ and $h > 0$, the discrete symmetric triangular kernel with rv $\mathcal{K}_{a;x,h}$ on support $\mathbb{S}_x = \{x-a, \ldots, x-1, x, x+1, \ldots, x+a\}$ has a pmf given by

$$\Pr(\mathcal{K}_{a;x,h} = z) = \frac{(a+1)^h - |y-x|^h}{P(a,h)}, \ z \in \mathbb{S}_x,$$

with $P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=1}^{a} k^h$ the normalizing constant. The modal probability and variance of this kernel can be developped as

$$\Pr(\mathcal{K}_{a;x,h} = x) = 1 - 2h\mathrm{A}(a) + O(h^2) \ \text{ and } \ \mathrm{Var}(\mathcal{K}_{a;x,h}) = 2h\mathrm{V}(a) + O(h^2),$$

with $\mathrm{A}(a) = a\log(a+1) - \sum_{k=1}^{a} \log(k)$ and $\mathrm{V}(a) = \{a(2a^2+3a+1)/6\}\log(a+1) - \sum_{k=1}^{a} k^2 \log(k)$. These expansions are useful in the following for establishing an expression of optimal bandwidth.

Concerning the data-driven bandwidth selection, an optimal bandwidth h_{opt} is obtained by minimizing the asymptotic part of mean integrated squared error (MISE) of $\widehat{m}_n$ in (2) using the discrete triangular kernel $K_{a;x,h}$ such that

$$\mathrm{MISE}\{\widehat{m}_n(x;a,h)\} = \mathrm{AMISE}\{\widehat{m}_n(x;a,h)\} + o\left(\frac{1}{n}\right) + O\left(h^2\right),$$

with the asymptotic MISE

$$\mathrm{AMISE}\{\widehat{m}_n(x;a,h)\} = \frac{h^2}{4}\mathrm{V}^2(a)\sum_{x\in\mathbb{T}} W^2(x) + \{1-h\mathrm{A}(a)\}^2 \sum_{x\in\mathbb{T}} \frac{\mathrm{Var}(Y|X=x)}{nf(x)}.$$

It ensues the optimal bandwidth

$$\widehat{h}_{opt}(a,n) = \frac{A(a)\sum_{x\in\mathbb{T}} \mathrm{Var}(Y|X=x)/f(x)}{\mathrm{A}^2(a)\sum_{x\in\mathbb{T}} \mathrm{Var}(Y|X=x)/f(x) + n\mathrm{V}^2(a)\sum_{x\in\mathbb{T}} W^2(x)} \sim C_0 n^{-1}$$

with constant C_0. Hence the root of $\mathrm{AMISE}\{\widehat{m}_n(x;a,h_{opt})\}$ is $O(n^{-1/2})$ and we get

$$m(x) = \widehat{m}_n(x;a,h_{opt}) + O(n^{-1/2}), \quad x \in \mathbb{T}.$$

3 Nonparametric kernel estimator for sensitivity analysis

This section aims at building the estimator of ANOVA decomposition of model f by using nonparametric kernel regression estimator. Let us first recall the ANOVA decomposition of $Y = f(X)$ given by

$$\begin{aligned} Y &= f_0 + \sum_{i=1}^{k} f_i(X_i) + \sum_{i<j} f_{ij}(X_i,X_j) + \ldots + f_{12\ldots k}(X_1,X_2,\ldots,X_k) \text{ with} \\ & f_0 = \mathbb{E}(Y), \quad f_i = \mathbb{E}(Y|X_i) - f_0, \quad f_{ij} = \mathbb{E}(Y|X_i,X_j) - f_i - f_j - f_0, \ldots \end{aligned} \tag{3}$$

It ensues the decomposition of the total variance of model output Y such as

$$\begin{aligned} \mathbb{V}(Y) &= \sum_{i=1}^{k} \mathbb{V}_i + \sum_{i<j} \mathbb{V}_{ij} + \ldots + \mathbb{V}_{12\ldots k}, \text{ with} \\ \mathbb{V}_i &= \mathbb{V}\{\mathbb{E}(Y|X_i)\}, \quad \mathbb{V}_{ij} = \mathbb{V}\{\mathbb{E}(Y|X_i,X_j)\} - \mathbb{V}_i - \mathbb{V}_j, \ldots \end{aligned} \tag{4}$$

3.1 Multivariate nonparametric regression

Let us consider $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{T}^d \subseteq \mathbb{N}^d$ a target vector and $\mathbf{H} = \mathbf{Diag}(h_{11}, \ldots, h_{dd})$ a bandwidth matrix with $h_{ii} > 0$ such that $\mathbf{H} \equiv \mathbf{H}_n$ goes to the null matrix $\mathbf{0_d}$ as $n \to \infty$. Assume $(\mathbf{X}^k, Y^k), k = 1, 2, \ldots, n$, be a sequence of iid random vectors defined on $\mathbb{T}^d \times \mathbb{R}$ with $m(\cdot) = \mathbb{E}(Y^k|\mathbf{X}^k = \cdot)$. The multivariate nonparametric regression estimator $\widehat{m}_n^d$ of m can be defined by

$$\widehat{m}_n^d(\mathbf{x};\mathbf{H}) = \sum_{k=1}^{n} \frac{Y^k K_{\mathbf{x},\mathbf{H}}(\mathbf{X}^k)}{\sum_{l=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{X}^l)}, \tag{5}$$

where the multivariate associated kernel $K_{x,\mathbf{H}}(\cdot) = \prod_{i=1}^{d} K^{[i]}_{x_i,h_{ii}}(\cdot)$ is defined as a product of univariate associated kernel $K^{[i]}_{x_i,h_{ii}}$ with its corresponding rv $\mathcal{K}^{[i]}_{x_i,h_{ii}}$ on support $\mathbb{S}_{x_i,h_{ii}}$, for all $i = 1, 2, \ldots, d$. Therefore, according to assumptions (A1), (A2) and (A3) for univariate associated kernel, the multivariate associated kernel of support $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{j=1}^{d} \mathbb{S}_{x_i,h_{ii}}$ is a pmf satisfying the following conditions:

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

where $\mathcal{K}_{\mathbf{x},\mathbf{H}}$ denotes the rv with pmf $K_{\mathbf{x},\mathbf{H}}$ and both $\mathbf{a}(\mathbf{x},\mathbf{H}) = (a_1(\mathbf{x},\mathbf{H}), \ldots, a_d(\mathbf{x},\mathbf{H}))^\top$ and $\mathbf{B}(\mathbf{x},\mathbf{H}) = (b_{ij}(\mathbf{x},\mathbf{H}))_{i,j=1,\ldots,d}$ tend, respectively, to null vector $\mathbf{0}$ and null matrix $\mathbf{0_d}$ as $\mathbf{H} \to \mathbf{0_d}$ (Sobom and Kokonendji, 2015).

For $\mathbf{a} = (a_1, a_2, \ldots, a_d)^\top \in \mathbb{N}^d$, the multivariate estimator $\widehat{m}^d_n$ using discrete symmetric triangular kernel $K_{\mathbf{a};\mathbf{x},\mathbf{H}}(\cdot) = \prod_{i=1}^{d} K^{[i]}_{a_i,x_i,h_{ii}}(\cdot)$ with an optimal bandwidth matrix $\mathbf{H}_{opt}$ such as $h_{opt,ii} \sim C_1 n^{-1/d}$ (with constant C_1) satisfies

$$m(\mathbf{x}) = \widehat{m}^d_n(\mathbf{x};\mathbf{a},\mathbf{H}_{opt}) + O(n^{-1/(d+1)}), \;\; \mathbf{x} \in \mathbb{T}^d.$$

3.2 Kernel estimator of ANOVA decomposition

Consider the ANOVA decomposition in (3) of the model f. The estimator of f_0 is the arithmetic average of Y^k, $\widehat{f_0} = 1/n \sum_{k=1}^{n} Y^k$. Then, the terms f_i are estimated by

$$\widehat{f_i}(x_i;h_{ii}) = \frac{1}{n}\sum_{k=1}^{n} K_{x_i,h_{ii}}(X^k_i)Y^k - \frac{1}{n}\sum_{k=1}^{n} Y^k = \frac{1}{n}\sum_{k=1}^{n} \mathbb{K}_{x_i,h_{ii}}(X^k_i)Y^k$$

with $\mathbb{K}_{x_i,h_{ii}}(X^k_i) = K_{x_i,h_{ii}}(X^k_i) - 1$. In the same way, the terms f_{ij} are estimated as follows:

$$\widehat{f_{ij}}(x_i,x_j;\mathbf{H}) = \widehat{\mathbb{E}}(Y^k|X^k_i,X^k_j) - \widehat{f_i} - \widehat{f_j} - \widehat{f_0} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{K}_{\mathbf{x},\mathbf{H}}(\mathbf{X}^k)Y^k$$

with $\mathbb{K}_{\mathbf{x},\mathbf{H}}(\cdot) = K_{\mathbf{x},\mathbf{H}}(\cdot) - K^{[i]}_{x_i,h_{ii}}(\cdot) - K^{[j]}_{x_j,h_{jj}}(\cdot) - 1$, where $K_{\mathbf{x},\mathbf{H}}(\cdot)$ is the multivariate associated kernel in (5). Thus, we get the estimation of the variance terms in (4):

$$\widehat{\mathbb{V}}(Y) = \mathbb{E}_{\mathbf{X}^k}\{\widehat{m}^d_n(\mathbf{x};\mathbf{H})\}^2 - \widehat{f_0}^2, \;\; \widehat{\mathbb{V}}_i = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_i}(x_i;h_{ii})\}^2, \;\; \widehat{\mathbb{V}_{ij}} = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_{ij}}(x_i,x_j;h_{ii},h_{jj})\}^2, \ldots$$

For example, it ensues the estimated Sobol first order indices in (1) given by

$$\widehat{S_i} = \frac{\widehat{\mathbb{V}}_i}{\widehat{\mathbb{V}}(Y)} = \frac{S_i + O(n^{-1/(d+1)})}{1 + O(n^{-1/(d+1)})} \to S_i \;\text{ as } n \text{ goes to } \infty.$$

Finally, the nonparametric estimator using discrete symmetric triangular kernel is applied in comparison to its continuous version using gaussian kernel on a test function analysis and on a real case study from agricultural (Andrianandraina *et al.*, 2014). By using Monte Carlo simulations, the discrete kernel estimator is revealed to be competing to continuous kernel estimation for estimating the Sobol SA indices of discrete input variables.

Bibliographie

[1] Abdous, B., Kokonendji, C.C. et Senga Kiessé, T. (2012), On semiparametric regression for count explanatory variables, *Journal of Statistical Planning and Inference*, 142, 1537–1548.

[2] Andrianandraina, Ventura, A., Senga Kiessé, T., Cazacliu,B., Rachida,I. et van der Werf, H.M.G. (2014), Action-oriented LCA: A novel approach to decision making in a life cycle context- a case study of hemp crop production, *Journal of Industrial Ecology* [accepted on july 2014].

[3] Iooss, B. (2011), Review of global sensitivity analysis of numerical models, *Journal de la Société Française de Statistique*, 152, 3–25.

[4] Kokonendji, C.C. et Senga Kiessé, T. (2011), Discrete associated kernel method for smoothing discrete function and extensions, *Statistical Methodology*, 8, 497–516.

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

[7] Ratto, M., Pagano, A. et Young, P. (2007), State Dependent Parameter metamodelling and sensitivity analysis, *Computer Physics Communications*, 177, 863-876.

[8] Rosenblatt, M. (1969), *Conditional probability density and regression estimates*, In:Krishnaiah PR, editor. Multivariate analysis, 2nd ed. p. 25–31.

[9] Sobol, I.M. (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation*, 55, 271–280.

[10] Senga Kiessé, T., Lorino, T. et Khraibani, H. (2014), Discrete nonparametric kernel and parametric methods for modeling pavement deterioration, *Communications in Statistics - Theory and Methods*, 43, 1164–1178.

[11] Sobom, M.S., Kokonendji, C.C. (2015), Effects of associated kernels in nonparametric multiple regressions, *http://arxiv.org/abs/1502.01488v1*.

[12] Watson, G. S. (1964), Smooth regression analysis, *Sankhyā Ser. A*, 26, 359–372.