Reliability Engineering and System Safety 146 (2016) 47–54

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Discrete non-parametric kernel estimation for global sensitivity analysis

Tristan Senga Kiessé *, Anne Ventura

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

ARTICLE INFO

Article history:
Received 19 February 2015
Received in revised form
12 September 2015
Accepted 12 October 2015
Available online 19 October 2015

Keywords:
Analysis of variance
Discrete kernel
Non-parametric regression
Sensitivity analysis
Sobol indice

ABSTRACT

This work investigates the discrete kernel approach for evaluating the contribution of the variance of discrete input variables to the variance of model output, via analysis of variance (ANOVA) decomposition. Until recently only the continuous kernel approach has been applied as a metamodeling approach within sensitivity analysis framework, for both discrete and continuous input variables. Now the discrete kernel estimation is known to be suitable for smoothing discrete functions. We present a discrete non-parametric kernel estimator of ANOVA decomposition of a given model. An estimator of sensitivity indices is also presented with its asymtotic convergence rate. Some simulations on a test function analysis and a real case study from agricultural have shown that the discrete kernel approach outperforms the continuous kernel one for evaluating the contribution of moderate or most influential discrete parameters to the model output.

1. Introduction

In the literature many works about reliability analysis approaches in general, and sensitivity analysis (SA) methods more specially, are related to different problems such as the important case of non-independent random inputs [6] and have various application domains such as maritime industry [19] or environment [14]. In most cases, a mathematical modeling of the studied system is frequently revealed to be useful when the variations of input parameters in a model imply a large variability of the results with some impacts on their accuracy. In this context, the probabilistic way is of interest to encompass the variation in the input parameters of the model. SA methods are then useful to conduct such a study since they aim to evaluate how the variation of input parameters contributes to the variation of the output of a model. Particularly, works in SA have highlighted the encountered interesting aspect concerning the evaluation of the influence of discrete (categorical or ordinal) inputs. Indeed, in system reliability studies, several models involving in various engineering contexts have input discrete variables. And, one of the reliability engineering issues is to accurately evaluate the influence of such parameters.

Amongst various SA approaches, let us consider a well-known method based on the analysis of variance (ANOVA) decomposition of model f for quantifying the influence of input $X_{i,i=1,2,\ldots,k} \in \mathbb{T}$ on the output $Y \in \mathbb{R}$. That method consists of the calculation of sensitivity indices given by [18] such that

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

The measure of first order S_i evaluates the contribution of the variation of X_i to the total variance of Y, the measure of second order S_{ij} evaluates the contribution of the interaction of X_i and X_j on the output, and so on. Various statistical tools as splines, generalized linear or additive model, polynomial are useful in a metamodeling approach for providing an estimation of conditional expectation $\mathbb{E}(Y|X_i)$ and, consequently, of the main effect sensitivity measure S_i [4]. In the framework of the non-parametric smoothing, some methods as the continuous kernel-based estimation [16] or the State-Dependent Parameter estimation [13] are good choices for estimating $\mathbb{E}(Y|X_i)$. About the two estimation methods, [15,20] are respectively one of the original references of nonparametric and state-dependent parameter estimates. Nowadays [11] have shown that continuous kernel estimation is equal or better than the SDP estimation in terms of performance. However until recently in the literature the continuous kernel estimation is evenly applied on continuous input variables as on discrete ones while discrete kernel estimation suitable for discrete functions is now known [7].

The discrete associated kernel method was developed for smoothing discrete functions as probability mass functions (pmf) or count regression functions on a discrete support $\mathbb{T}$ such as $\mathbb{T} = \mathbb{N}$, the set of positive integers, or $\mathbb{T} = \mathbb{Z}$, the set of integers. For

* Corresponding author. Tel.: +33 2 72 64 87 40.
E-mail addresses: tristan.sengakiesse@univ-nantes.fr (T. Senga Kiessé), anne.ventura@univ-nantes.fr (A. Ventura).

http://dx.doi.org/10.1016/j.ress.2015.10.010
0951-8320/

a fixed target x on discrete support $\mathbb{T}$ and a smoothing parameter $h>0$, this method is based on the definition of the *discrete associated kernel* $K_{x,h}(\cdot)$ which is a pmf of random variable (rv) $\mathcal{K}_{x,h}$ with support $\mathbb{S}_x$ satisfying

$x \in \mathbb{S}_x \quad (A1)$,

$\lim_{h\to 0}\mathbb{E}(\mathcal{K}_{x,h}) = x \quad (A2)$,

$\lim_{h\to 0}\mathbb{V}(\mathcal{K}_{x,h}) = 0 \quad (A3)$.

These three assumptions, fulfilled by both continuous and discrete kernels, insure good asymptotic properties for the corresponding kernel estimator [10]. Thus, for $(a,x)\in\mathbb{N}\times\mathbb{T}$ and $h>0$, an example of discrete associated kernel is the discrete symmetric triangular one with rv $\mathcal{K}_{a;x,h}$ on support $\mathbb{S}_x=\{x-a,\ldots,x-1,x,x+1,\ldots,x+a\}$ with 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 $P(a,h)=(2a+1)(a+1)^h-2\sum_{k=1}^{a}k^h$ a normalizing constant. From the discrete kernel methodology, a discrete non-parametric estimator of $\mathbb{E}(Y|X_i)$ was proposed by [1] adapted from the continuous version of [12,22] as follows:

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

with the arbitrary sequence of smoothing parameters $h=h(n)>0$ fulfilling $\lim_{n\to\infty}h(n)=0$ and $K_{x,h}(\cdot)$ a discrete associated kernel as defined previously.

In this paper the non-parametric regression estimator $\widehat{m}_n$ using a discrete symmeric triangular kernel is investigated as a novel approach in SA methods for providing estimated sensitivity indices for discrete input variables X_i. Thus, the discrete kernel estimation approach is studied as a contribution to reliability analysis for model with discrete input parameters. To illustrate the performance of discrete kernel approach in comparison to continuous kernel approach, some simulations are realized using Ishigami test function and an application is proposed on a real case from agricultural. That latter concerns the evaluation of the influence of some parameters on the environmental impacts generated during the Hemp Crop production by farmers [2].

2. Non-parametric discrete triangular regression

This section presents first a review of the non-parametric univariate regression estimator using symmetric discrete triangular kernel with the asymptotic expansion of its global squared error as presented by [3]. Herein, the optimal convergence rate of the discrete triangular regression estimator is added.

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 non-parametric 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. For $a\in\mathbb{N}$, a fixed point $x\in\mathbb{T}$ and a smoothing parameter $h>0$, let us consider the discrete non-parametric estimator $\widehat{m}_n$ of m defined in (2) using a discrete triangular symmetric kernel such that

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

First, about some asymptotic properties of estimator $\widehat{m}_n(a;x,h)$ in (2), the asymptotic part of its mean integrated squared error MISE [21] defined by

$$\mathrm{MISE}\{\widehat{m}_n(x;a,h)\}=\sum_{x\in\mathbb{T}}\mathrm{Var}\{\widehat{m}_n(x;a,h)\}+\sum_{x\in\mathbb{T}}\mathrm{Bias}^2\{\widehat{m}_n(x;a,h)\}.$$

is given by

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

This last expression is obtained by calculating asymptotic bias and variance of $\widehat{m}_n(x;a,h)$ in (2) using the following expansions of the modal probability and variance of the discrete symmetric triangular kernel:

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

with $\mathrm{A}(a)=a\log(a+1)-\sum_{k=1}^{a}\log(k)$ and $\mathrm{V}(a)=\{a(2a^2+3a+1)/6\}\log(a+1)-\sum_{k=1}^{a}k^2\log(k)$ (refer to [3] for more details). Then, an asymptotical optimal bandwidth h_{opt} is obtained by minimizing the asymptotic part AMISE of $\widehat{m}_n(a;x,h)$ in (2) such that

$$\widehat{h}_{opt}(a,n)=\frac{A(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{A}^2(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)+n\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)}\sim C_0n^{-1}$$

with

$$C_0=\frac{\mathrm{A}(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)}.$$

Finally, we get the following inequality:

$$\begin{aligned}\mathrm{AMISE}\{\widehat{m}_n(x;a,h_{opt})\}&\sim n^{-1}\left[\frac{C_0^2}{n}\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)\right.\\&\left.+\left\{1-\frac{C_0}{n}A(a)\right\}^2\sum_{x\in\mathbb{T}}\frac{\mathrm{Var}(Y|X=x)}{f(x)}\right]\le n^{-1}\left(C_0^2\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)\right.\\&\left.+\left[1+\{C_0A(a)\}^2\right]\sum_{x\in\mathbb{T}}\frac{\mathrm{Var}(Y|X=x)}{f(x)}\right)\end{aligned}$$

where AMISE $\{\widehat{m}_n(x;a,h_{opt})\}$ tends to 0 as $n\to\infty$. Thus, for $a\in\mathbb{N}$, the optimal asymptotic root MISE of estimator $\widehat{m}_n$ with kernel $K_{a;x,h}$ is $O(n^{-1/2})$ resulting in

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

Note that the discrete kernel estimation and the resulting asymptotic expansions of estimator's bias and variance depend on two preconditions: discrete random variable and smooth hypothesis. For $x\in\mathbb{T}$, a discrete associated kernel satisfying assumptions (A1)–(A3) has asymptotically the same behavior that a Dirac type kernel $D_x(y), y\in\mathbb{S}_x$, such that $D_x(y)=1$ at $y=x$ and 0 for any $y\neq x$. That explains also the good asymptotic properties of the corresponding estimator.

3. Non-parametric kernel estimator for sensitivity analysis

This section aims at building the estimator of ANOVA decomposition of the model $Y=f(X_1,X_2,\ldots,X_k)$ given by

$$Y=f_0+\sum_{i=1}^{k}f_i(X_i)+\sum_{i<j}f_{ij}(X_i,X_j)+\cdots+f_{12\ldots k}(X_1,X_2,\ldots,X_k), \tag{3}$$

where each term is defined by

$$f_0=\mathbb{E}(Y), f_i=\mathbb{E}(Y|X_i)-f_0,\quad f_{ij}=\mathbb{E}(Y|X_i,X_j)-f_i-f_j-f_0,\ldots \tag{4}$$

Non-parametric kernel estimation of such model originates in the work of [11] for continuous case. The multidimensional version of non-parametric regression estimator $\widehat{m}_n$ is presented for the calculation of Sobol indices when measuring the contribution of two or more variables to the variance of Y.