Reliability Engineering and System Safety 112 (2013) 67–81

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

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

Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes

Samir Touzani *, Daniel Busby

IFP Energies nouvelles, 92852 Rueil-Malmaison, France

ARTICLE INFO

Article history:
Received 7 September 2011
Received in revised form
6 November 2012
Accepted 7 November 2012
Available online 2 December 2012

Keywords:
Global sensitivity analysis
Metamodel
Smoothing spline ANOVA
Nonparametric regression

ABSTRACT

The paper investigates a nonparametric regression method based on smoothing spline analysis of variance (ANOVA) approach to address the problem of global sensitivity analysis (GSA) of complex and computationally demanding computer codes. The two steps algorithm of this method involves an estimation procedure and a variable selection. The latter can become computationally demanding when dealing with high dimensional problems. Thus, we proposed a new algorithm based on Landweber iterations. Using the fact that the considered regression method is based on ANOVA decomposition, we introduced a new direct method for computing sensitivity indices. Numerical tests performed on several analytical examples and on an application from petroleum reservoir engineering showed that the method gives competitive results compared to a more standard Gaussian process approach.

1. Introduction

The recent significant advances in computational power have allowed computer modeling and simulation to become an integral tool in many industrial and scientific applications, such as nuclear safety assessment, meteorology or oil reservoir forecasting. Simulations are performed with complex computer codes that model diverse complex real world phenomena. Inputs of such computer codes are estimated by experts or by indirect measures and can be highly uncertain. It is important to identify the most significant inputs, which contribute to the model prediction variability. This task is generally performed by the variance-based sensitivity analysis also known as global sensitivity analysis (GSA) (see [1] and [2]).

The aim of GSA for computer codes is to quantify how the variation of an output of the computer code is apportioned to different inputs of the model. The most useful methods that perform sensitivity analysis require stochastic simulation techniques, such as Monte-Carlo methods. These methods usually involve several thousands computer code evaluations that are generally not affordable with realistic models for which each simulation requires several minutes, hours or days. Consequently, meta-modeling methods become an interesting alternative.

A meta-model is an approximation of the computer code's input/output relation, which is fast to evaluate. The general idea of this approach is to perform a limited number of model evaluations (hundreds) at some carefully chosen training input values, and then, using statistical regression techniques to construct an approximation of the model. If the resulting approximation is of a good quality, the meta-model is used instead of the complex and computationally demanding computer code to perform the GSA.

The most commonly used meta-modeling methods are those based on parametric polynomial regression models, which require specifying the polynomial form of the regression mean (linear, quadratic, etc.). However, it is often the case that the linear (or quadratic) model can fail to identify properly the input/output relation. Thus, in nonlinear situations, nonparametric regression methods are preferred.

In the last decade many different nonparametric regression models have been used as a meta-modeling method. To name a few of them [3–5] utilized a Gaussian Process (GP). [6,7] used a polynomial chaos expansions to perform a GSA.

In addition [8–10] provide a comparison of various parametric and nonparametric regression models, such as linear regression (LREG), quadratic regression (QREG), projection pursuit regression multivariate adaptive regression splines (MARS), gradient boosting regression, random forest, Gaussian process (GP), adaptive component selection and smoothing operator (ACOSSO), etc… for providing appropriate metamodel strategies.

We focus in this work on the modern nonparametric regression method based on smoothing spline ANOVA (SS-ANOVA) model and component selection and smoothing operator (COSSO) regularization, which can be seen as an extension of the LASSO

* Corresponding author.
E-mail addresses: samirtouzani.phd@gmail.com (S. Touzani), daniel.busby@ifpen.fr (D. Busby).

0951-8320/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ress.2012.11.008

[11] variable selection method in parametric models to nonparametric models. Moreover, we use the ANOVA decomposition basis of the COSSO to introduce a direct method to compute the sensitivity indices.

In this paper, we first review the SS-ANOVA, then we will describe the COSSO method and its algorithm. Furthermore we will introduce two new algorithms which provide the COSSO estimates, the first one using an iterative algorithm based on Landweber iterations and the second one using a modified least angle regression algorithm (LARS) (see [12,13]). Next we will describe our new method to compute the sensitivity indices. Finally, numerical simulations and an application from petroleum reservoir engineering will be presented and discussed.

2. Smoothing spline ANOVA approach for metamodels

In mathematical terms, the computer code can be represented as a function $Y=f(\mathbf{X})$ where Y is the output scalar of the computer code, $\mathbf{X}=(X^{(1)},\dots,X^{(d)})$ the d-dimensional inputs vector which represent the uncertain parameters and $f:\mathbb{R}^d\rightarrow\mathbb{R}$ is the unknown function that represents the computer code. Our purpose is to introduce an estimation procedure for f.

A popular approach to the nonparametric estimation for high dimensional problems is the smoothing spline analysis of variance (SS-ANOVA) model [14]. To remind, the ANOVA expansion is defined as

$$f(\mathbf{X})=f_0+\sum_{j=1}^{d} f_j(X^{(j)})+\sum_{j<l} f_{jl}(X^{(j)},X^{(l)})+\cdots+f_{1,\dots,d}(X^{(1)},\dots,X^{(d)}) \tag{1}$$

where f_0 is a constant, f_j's are univariate functions representing the main effects, f_{jl}'s are bivariate functions representing the two way interactions, and so on.

It is important to determine which ANOVA components should be included in the model. Lin and Zhang [15] proposed a penalized least square method with the penalty functional being the sum of component norms. The COSSO is a regularized nonparametric regression method based on ANOVA decomposition.

In the following subsections, we first review the definition of the COSSO method. Then we present the COSSO algorithm proposed in [15]. Finally we introduce three new variations of the COSSO algorithm.

2.1. Definition

Let $f\in\mathcal{F}$, where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) (for more details we refer to [14,16]) corresponding to the ANOVA decomposition (1), and let $\mathcal{H}^j=\{1\}\oplus\overline{\mathcal{H}}^j$ be a RKHS of functions of $X^{(j)}$ over [0,1], where {1} is the RKHS consisting of only the constant functions and $\overline{\mathcal{H}}^j$ is the RKHS consisting of functions $f_j\in\mathcal{H}^j$ such that $\langle f_j,1\rangle_{\mathcal{H}^j}=0$. Then the model space $\mathcal{F}$ is the tensor product space of $\mathcal{H}^j$

$$\mathcal{F}=\bigotimes_{j=1}^{d}\mathcal{H}^j=\{1\}\oplus\sum_{j=1}^{d}\overline{\mathcal{H}}^j\oplus\sum_{j<l}[\overline{\mathcal{H}}^j\otimes\overline{\mathcal{H}}^l]\dots \tag{2}$$

Each component in the ANOVA decomposition (1) is associated to a corresponding subspace in the orthogonal decomposition (2). We assume that only second order interactions are considered in the ANOVA decomposition and an expansion to the second order generally provides a satisfactory description of the model.

Let us consider the index $\alpha\equiv j$ for $\alpha=1,\dots,d$ with $j=1,\dots,d$ and $\alpha\equiv(j,l)$ for $\alpha=d+1,\dots,d(d+1)/2$ (where $d(d+1)/2$ corresponds to the number of ANOVA components) with $1\le j<l\le d$. Using such notation in (2) we obtain

$$\mathcal{F}=\{1\}\oplus\bigoplus_{\alpha=1}^{q}\mathcal{F}^\alpha=\{1\}\oplus\sum_{j=1}^{d}\overline{\mathcal{H}}^j\oplus\sum_{j<l}[\overline{\mathcal{H}}^j\otimes\overline{\mathcal{H}}^l] \tag{3}$$

where $\mathcal{F}^1,\dots\mathcal{F}^q$ are q orthogonal subspaces of $\mathcal{F}$ and $q=d(d+1)/2$. We denote by $\|\cdot\|$ the norm in the RKHS $\mathcal{F}$. For some λ the COSSO estimate is given by the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-f(\mathbf{x}_i))^2+\lambda^2\sum_{\alpha=1}^{q}\|P_\alpha f\| \tag{4}$$

where λ is the regularization parameter and P_α is the orthogonal projection onto $\mathcal{F}^\alpha$.

2.2. COSSO algorithm

Lin and Zhang [15] have shown that the minimizer of (4) has the form $\hat{f}=\hat{b}+\sum_{\alpha=1}^{q}\hat{f}_\alpha$, with $\hat{f}_\alpha\in\mathcal{F}^\alpha$. By the reproducing kernel property of $\mathcal{F}^\alpha$, $\hat{f}_\alpha\in\text{span}\{K_\alpha(x_i,\cdot),\ i=1,\dots,n\}$, where K_α is the reproducing kernel of $\mathcal{F}^\alpha$ defined by

$$K_\alpha(x,x')=K_j(x,x')=k_1(x)k_1(x')+k_2(x)k_2(x')-k_4(|x-x'|)$$

where $k_l(x)=B_l(x)/l!$ and B_l is the lth Bernoulli polynomial. Thus, for $x\in[0,1]$

$$k_1(x)=x-\frac{1}{2}$$

$$k_2(x)=\frac{1}{2}(k_1^2(x)-1/12)$$

$$k_4(x)=\frac{1}{24}\left(k_1^4(x)-\frac{k_1^2(x)}{2}+\frac{7}{240}\right)$$

Moreover, the reproducing kernel K_α for the RKHS $\mathcal{F}^\alpha$ such as $\mathcal{F}^\alpha\equiv\overline{\mathcal{H}}^j\otimes\overline{\mathcal{H}}^l$, are given by the following tensor products:

$$K_\alpha(\mathbf{X},\mathbf{X}')=K_j(X^{(j)},X^{(j)'})K_l(X^{(l)},X^{(l)'})$$

For more details we refer to [14].

Lin and Zhang [15] have also shown that (4) is equivalent to a more easier form to compute, which is

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i-f(\mathbf{x}_i)\}^2+\lambda_0\sum_{\alpha=1}^{q}\theta_\alpha^{-1}\|P_\alpha f\|^2+\nu\sum_{\alpha=1}^{q}\theta_\alpha \text{ subject to } \theta_\alpha\ge 0 \tag{5}$$

where λ_0 is a constant and ν is a smoothing parameter. If $\theta_\alpha=0$, then the minimizer of (5) is taken to satisfy $\|P_\alpha f\|=0$ and we use the convention $0/0=0$. The penalty term of θ's, $\sum_{\alpha=1}^{q}\theta_\alpha$, controls the sparsity of each component f_α.

For fixed $\boldsymbol{\theta}=(\theta_1,\dots,\theta_q)^T$ (5) is equivalent to the standard SS-ANOVA [14] and therefore the solution has the form

$$f(\mathbf{x})=b+\sum_{i=1}^{n}c_i\sum_{\alpha=1}^{q}\theta_\alpha K_\alpha(\mathbf{x}_i,\mathbf{x}) \tag{6}$$

and let K_α be the $n\times n$ matrix $\{K_\alpha(\mathbf{x}_i,\mathbf{x}_j)\}$, $i=1,\dots,n$, $j=1,\dots,n$, let K_θ stand for the matrix $\sum_{\alpha=1}^{q}\theta_\alpha K_\alpha$ and $\mathbf{1}_n$ the column vector consisting of n ones. Then $f=K_\theta\mathbf{c}+b\mathbf{1}_n$ with $\mathbf{c}=(c_1,\dots,c_n)^T$ and (5) can be expressed as

$$\frac{1}{n}\left\|\mathbf{Y}-\sum_{\alpha=1}^{q}\theta_\alpha K_\alpha\mathbf{c}-b\mathbf{1}_n\right\|^2+\lambda_0\mathbf{c}^TK_\theta\mathbf{c}+\nu\sum_{\alpha=1}^{q}\theta_\alpha \tag{7}$$

For a fixed $\boldsymbol{\theta}$, (7) can be written as

$$\min_{\mathbf{c},b}\|\mathbf{Y}-K_\theta\mathbf{c}-b\mathbf{1}_n\|^2+n\lambda_0\mathbf{c}^TK_\theta\mathbf{c} \tag{8}$$