Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Calculations of Sobol indices for the Gaussian process metamodel

Amandine Marrel [a,*], Bertrand Iooss [b], Béatrice Laurent [c], Olivier Roustant [d]

[a] CEA, DEN, DTN/SMTM/LMTE, F-13108 Saint Paul lez Durance, France
[b] CEA, DEN, DER/SESI/LCFR, F-13108 Saint Paul lez Durance, France
[c] Institut de Mathématiques, Université de Toulouse (UMR 5219), France
[d] Ecole des Mines de Saint-Etienne, France

ARTICLE INFO

Article history:
Received 30 October 2007
Received in revised form
7 April 2008
Accepted 12 July 2008
Available online 23 July 2008

Keywords:
Gaussian process
Covariance
Metamodel
Sensitivity analysis
Uncertainty
Computer code

ABSTRACT

Global sensitivity analysis of complex numerical models can be performed by calculating variance-based importance measures of the input variables, such as the Sobol indices. However, these techniques, requiring a large number of model evaluations, are often unacceptable for time expensive computer codes. A well-known and widely used decision consists in replacing the computer code by a metamodel, predicting the model responses with a negligible computation time and rending straightforward the estimation of Sobol indices. In this paper, we discuss about the Gaussian process model which gives analytical expressions of Sobol indices. Two approaches are studied to compute the Sobol indices: the first based on the predictor of the Gaussian process model and the second based on the global stochastic process model. Comparisons between the two estimates, made on analytical examples, show the superiority of the second approach in terms of convergence and robustness. Moreover, the second approach allows to integrate the modeling error of the Gaussian process model by directly giving some confidence intervals on the Sobol indices. These techniques are finally applied to a real case of hydrogeological modeling.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Environmental risk assessment is often based on complex computer codes, simulating for instance an atmospheric or hydrogeological pollution transport. These computer models calculate several output values (scalars or functions) which can depend on a high number of input parameters and physical variables. To provide guidance to a better understanding of this kind of modeling and in order to reduce the response uncertainties most effectively, sensitivity measures of the input importance on the response variability can be useful [1–6]. However, the estimation of these measures (based on Monte-Carlo methods for example) requires a large number of model evaluations, which is untractable for time expensive computer codes. Solutions exist to reduce this computational cost like random balance designs [7] but, in our industrial problems, the choice of the sampling design is not always available. Moreover, we want to use a generic tool to make both sensitivity analysis and uncertainty propagation and, more generally, any prediction of computer code. This kind of problem is of course not limited to environmental modeling and can be applied to any simulation system.

To overcome the problem of huge calculation time in sensitivity analysis, approaches based on nonparametric estimation tools have been proposed by Doksum and Samarov [8] and more recently by Da Veiga and Gamboa [9]. These nonparametric methods allow to significantly reduce the number of function evaluations needed to accurately estimate sensitivity indices. Another solution that we want to focus on in this paper can be to replace the complex computer code by a mathematical approximation, called a response surface or a surrogate model or also a metamodel. The response surface method [10] consists in constructing a function from few experiments that simulate the behavior of the real phenomenon in the domain of influential parameters. These methods have been generalized to develop surrogates for costly computer codes [11,12]. Several metamodels are classically used: polynomials, splines, generalized linear models, or learning statistical models like neural networks, regression trees, support vector machines [13,14]. Besides, Ratto et al. [15] recently proposed to use a State Dependent Parameter metamodeling to build an approximation of the computer code and perform the sensitivity analysis studies.

Among all the solutions based on metamodels, our attention is focused on the Gaussian process model which can be viewed as an extension of the kriging principles [11,16,17]. This metamodel which is characterized by its mean and covariance functions, presents several advantages: it is an exact interpolator and it is

[*] Corresponding author. Tel.: +33 4 42 25 26 52; fax: +33 4 42 25 62 72.
E-mail address: amandine.marrel@cea.fr (A. Marrel).

0951-8320/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2008.07.008

interpretable (not a black-box function). Moreover, numerous authors (for example, [6,18–20]) have shown how this model can provide a statistical basis for computing an efficient predictor of code response. In addition to its efficiency, this model gives an analytical formula which is very useful for sensitivity analysis, especially for the variance-based importance measures, the so-called Sobol indices [1,2]. To derive analytical expression of Sobol indices, Chen et al. [21] used tensor-product formulation and Oakley and O'Hagan [22] considered the Bayesian formalism of Gaussian processes.

We propose to compare these two analytical formulations of Sobol indices for the Gaussian process model: the first is obtained considering only the predictor, i.e. the mean of the Gaussian process model [21], while the second is obtained using all the global stochastic model [22]. In the last case, the estimate of a Sobol index is itself a random variable. Its standard deviation is available and we propose an original algorithm to estimate its distribution. Consequently, our method leads to build confidence intervals for the Sobol indices. To our knowledge, this information has not been proposed before and can be obtained thanks to the analytical formulation of the Gaussian process model error. This is particularly interesting in practice when the predictive quality of the metamodel is not high (because of small learning sample size for example), and our confidence on Sobol index estimates via the metamodel is poor.

The next section briefly explains the Gaussian process modeling and the Sobol indices defined in the two approaches (predictor-only and global model). In Section 3, the numerical computation of a Sobol index is presented. In the case of the global stochastic model, a procedure is developed to simulate its distribution. Section 4 is devoted to applications on analytical functions. First, comparisons are made between the Sobol indices based on the predictor and those based on the global model. The pertinence of simulating all the distribution of Sobol indices is therefore evaluated. Finally, Sobol indices and their uncertainty are computed for a real data set coming from a hydrogeological transport model based on waterflow and diffusion dispersion equations. The last section provides some possible extensions and concluding remarks.

2. Sobol indices with Gaussian process model

2.1. Gaussian process model

Let us consider n realizations of a computer code. Each realization $y(\boldsymbol{x})$ of the computer code output corresponds to a d-dimensional input vector $\boldsymbol{x} = (x_1, \dots, x_d)$. The n input points corresponding to the code runs are called an experimental design and are denoted as $X_s = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)})$. The outputs will be denoted as $Y_s = (y^{(1)}, \dots, y^{(n)})$ with $y^{(i)} = y(\boldsymbol{x}^{(i)}), i = 1, \dots, n$. Gaussian process (Gp) modeling treats the deterministic response $y(\boldsymbol{x})$ as a realization of a random function $Y(\boldsymbol{x})$, including a regression part and a centered stochastic process. The sample space Ω denotes the space of all possible outcomes ω, which is usually the Lebesgue-measurable set of real numbers. The Gp is defined on $R^d \times \Omega$ and can be written as

$$Y(\boldsymbol{x}, \omega) = f(\boldsymbol{x}) + Z(\boldsymbol{x}, \omega). \tag{1}$$

In the following, we use indifferently the terms Gp model and Gp metamodel.

The deterministic function $f(\boldsymbol{x})$ provides the mean approximation of the computer code. Our study is limited to the parametric case where the function f is a linear combination of elementary functions. Under this assumption, $f(\boldsymbol{x})$ can be written as follows:

$$f(\boldsymbol{x}) = \sum_{j=0}^{k} \beta_j f_j(\boldsymbol{x}) = F(\boldsymbol{x})\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = [\beta_0, \dots, \beta_k]^{\mathrm{t}}$ is the regression parameter vector, f_j $(j = 1, \dots, k)$ are basis functions and $F(\boldsymbol{x}) = [f_0(\boldsymbol{x}), \dots, f_k(\boldsymbol{x})]$ is the corresponding regression vector. In the case of the one-degree polynomial regression, $(d+1)$ basis functions are used:

$$\begin{cases} f_0(\boldsymbol{x}) = 1, \\ f_i(\boldsymbol{x}) = x_i \quad \text{for } i = 1, \dots, d. \end{cases}$$

In our applications, we use this one-degree polynomial as the regression part in order to simplify all the analytical numerical computation of sensitivity indices. This can be extended to other bases of regression functions. Without prior information on the relationship between the output and the input, a basis of one-dimensional functions is recommended to simplify the computations in sensitivity analysis and to keep one of the most advantages of Gp model [23].

The stochastic part $Z(\boldsymbol{x}, \omega)$ is a Gaussian centered process fully characterized by its covariance function: $\mathrm{Cov}_\Omega(Z(\boldsymbol{x}, \omega), Z(\boldsymbol{u}, \omega)) = \sigma^2 R(\boldsymbol{x}, \boldsymbol{u})$, where σ^2 denotes the variance of Z and R is the correlation function that provides interpolation and spatial correlation properties. To simplify, a stationary process $(Z(\boldsymbol{x}, \omega))$ is considered, which means that the correlation between $Z(\boldsymbol{x}, \omega)$ and $Z(\boldsymbol{u}, \omega)$ is a function of the difference between $\boldsymbol{x}$ and $\boldsymbol{u}$. Moreover, our study is restricted to a family of correlation functions that can be written as a product of one-dimensional correlation functions:

$$\mathrm{Cov}_\Omega(Z(\boldsymbol{x}, \omega), Z(\boldsymbol{u}, \omega)) = \sigma^2 R(\boldsymbol{x} - \boldsymbol{u}) = \sigma^2 \prod_{l=1}^{d} R_l(x_l - u_l). \tag{2}$$

This form of correlation function is particularly well adapted to get some simplifications of the integrals in the future analytical developments: in the case of independent inputs, it implies the computation of only one or two-dimensional integrals to compute the Sobol indices. Indeed, as described in Section 3.2, the application and the computation of the Sobol index formulae are simplified when the correlation function has the form of a one-dimensional product [6].

Among other authors, Chilès and Delfiner [24] and Rasmussen and Williams [20] give a list of correlation functions with their advantages and drawbacks. Among all these functions, our attention is devoted to the generalized exponential correlation function:

$$R_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x} - \boldsymbol{u}) = \prod_{l=1}^{d} \exp(-\theta_l |x_l - u_l|^{p_l}) \quad \text{with } \theta_l \geqslant 0 \text{ and } 0 < p_l \leqslant 2,$$

where $\boldsymbol{\theta} = [\theta_1, \dots, \theta_d]^{\mathrm{t}}$ and $\boldsymbol{p} = [p_1, \dots, p_d]^{\mathrm{t}}$ are the correlation parameters. This choice is motivated by the derivation and regularity properties of this function. Moreover, within the range of covariance parameters values, a wide spectrum of shapes is possible: for example, $p = 1$ gives the exponential correlation function while $p = 2$ gives the Gaussian correlation function.

2.2. Joint and conditional distributions

Under the hypothesis of a Gp model, the learning sample Y_s follows a multivariate normal distribution $p_\Omega(Y_s | X_s)$:

$$p_\Omega(Y_s, \omega | X_s) = \mathcal{N}(F_s \boldsymbol{\beta}, \boldsymbol{\Sigma}_s),$$

where $F_s = [F(\boldsymbol{x}^{(1)})^{\mathrm{t}}, \dots, F(\boldsymbol{x}^{(n)\mathrm{t}})]$ is the regression matrix and

$$\boldsymbol{\Sigma}_s = \sigma^2 R_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})_{i,j=1\dots n}$$

is the covariance matrix.

Download English Version:

https://daneshyari.com/en/article/807000

Download Persian Version:

https://daneshyari.com/article/807000

Daneshyari.com