Reliability Engineering and System Safety 96 (2011) 440–449

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

The identification of model effective dimensions using global sensitivity analysis

Sergei Kucherenko [a,*], Balazs Feil [b], Nilay Shah [a], Wolfgang Mauntz [c]

[a] CPSE, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] Department of Process Engineering, University of Pannonia, Veszprem, Hungary
[c] Lehrstuhl für Anlagensteuerungstechnik, Fachbereich Chemietechnik, Universität Dortmund, Germany

ARTICLE INFO

Article history:
Received 23 January 2010
Received in revised form
9 November 2010
Accepted 18 November 2010
Available online 25 November 2010

Keywords:
Monte Carlo integration
Quasi-Monte Carlo
Global sensitivity analysis
Sobol' sensitivity indices
Effective dimensions

ABSTRACT

It is shown that the effective dimensions can be estimated at reasonable computational costs using variance based global sensitivity analysis. Namely, the effective dimension in the truncation sense can be found by using the Sobol' sensitivity indices for subsets of variables. The effective dimension in the superposition sense can be estimated by using the first order effects and the total Sobol' sensitivity indices. The classification of some important classes of integrable functions based on their effective dimension is proposed. It is shown that it can be used for the prediction of the QMC efficiency. Results of numerical tests verify the prediction of the developed techniques.

1. Introduction

Modern mathematical models of real systems in physics, chemistry, biology, economics and other areas often have high complexity with hundreds or even thousands of variables. Straightforward modelling using such models can be computationally costly or even impossible. There is a demand for complexity reduction techniques which are not only general and applicable to any complex non-linear model but also rigorous in that their application provides estimates of the approximation errors. Variance based global sensitivity analysis allows to develop such complexity reduction techniques. Recently a new class of measures was introduced by Borgonovo [1,2]. These measures are known as moment-independent. They are based on the entire distribution of the output without a specific reference to its moments. Potentially, moment-independent measures can also be used for complexity reduction.

For modelling and complexity reduction purposes it is important to distinguish between the model nominal dimension and its effective dimension. The notions of the "effective dimension" in the truncation and superposition sense was introduced by Caflisch et al. in [3]. Quite often complex mathematical models have effective dimensions much lower than their nominal dimensions. The knowledge of model effective dimensions is very important as it allows to apply various complexity reduction techniques.

The effective dimension in the truncation sense d_T loosely speaking is equal to the number of important factors in the model. Identification of important and not important variables allows to fix not important variables at their nominal values. The resultant model would have lower complexity with dimensionality reduced from n to d_T. A condition $d_T \ll n$ often occurs in practical problems. Another type of complexity reduction is associated with the effective dimension in the superposition sense d_S: the function has the effective dimension in the superposition sense d_S if it is almost a sum of s-dimensional function components in the ANOVA decomposition.

For some problems such as path-dependent option pricing in mathematical finance changing the order in which input variables are sampled can dramatically decrease d_T. Such techniques are known as dimension reduction. Most results on dimension reduction are empirical and qualitative (see for example [3]).

A straightforward evaluation of the effective dimensions from their definitions is not practical in the general. Owen introduced the dimension distribution for a square integrable function [4]. The effective dimension can be defined through a quantile of the dimension distribution. He showed that for some classes of functions quantiles of the dimension distribution can be explicitly calculated but they are difficult to estimate in a general case. In this paper we show that global sensitivity analysis based on the Sobol' sensitivity indices (SI) allows to estimate the effective dimensions at reasonable computational costs.

Evaluation of the Sobol' SI necessitates the computation of high-dimensional integrals. The classical grid methods become computationally impractical when the number of dimensions n increases

* Corresponding author. Tel.: +44 207 594 6623; fax: +44 207 594 6606.
E-mail address: s.kucherenko@ic.ac.uk (S. Kucherenko).

0951-8320/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2010.11.003

because of "the curse of dimensionality". The convergence rate of Monte Carlo (MC) integration methods does not depend on the number of dimensions n. However, the rate of convergence $O(N^{-1/2})$, where N is the number of sampled points, attained by MC is rather slow. A higher rate of convergence can be obtained by using quasi-Monte Carlo (QMC) methods based on uniformly distributed sequences instead of pseudo-random numbers. Asymptotically, QMC can provide the rate of convergence $O(N^{-1})$.

For sufficiently large N, QMC should always outperform MC. However, in practice such sample sizes quite often are infeasible, especially when high-dimensional problems are concerned. Many numerical experiments demonstrated that the advantages of QMC can disappear for high-dimensional problems. There were claims that the degradation in performance of QMC occurs at $n \geq 12$ [5]. In contrast, other papers reported the superiority of QMC over MC for some integrands with $n = 360$ [6]. Some explanations for such inconsistent results were given using the notion of the effective dimension [3]. In [7] it was shown how the ANOVA components are linked to the effectiveness of QMC integration methods. Sloan and Wozniakowski [8] studied the efficiency of the quasi-Monte Carlo algorithms for high-dimensional integrals. They identified classes of functions for which the effect of the dimension is negligible. These are the so-called weighted classes in which the behavior in the successive dimensions is moderated by a sequence of weights.

There is no computationally feasible technique that would predict the efficiency of QMC in high dimensions. In this paper we use Sobol' SI as a quantitative measure of the QMC efficiency.

This paper is organized as follows. Section 2 briefly describes MC and QMC integration algorithms and issues concerning the possible degradation of QMC efficiency in higher dimensions. Section 3 gives a description of the Sobol' SI. Section 4 presents improved formulas for evaluation of the Sobol' SI. The notion of the effective dimension is introduced in Section 5. The classification of functions based on Sobol' SI is suggested in Section 6. It is shown how this classification can be used for the prediction of the QMC efficiency. Test examples and numerical results are considered in Section 7. Finally, conclusions are given in Section 8.

2. MC and QMC algorithms

Consider the evaluation of an integral

$$I[f] = \int_{H^n} f(\mathbf{x})\, d\mathbf{x},$$

where the function $f(\mathbf{x})$ is integrable in the n-dimensional unit hypercube H^n and sufficiently regular. The Monte Carlo quadrature formula is based on the probabilistic interpretation of an integral. An approximation to this expectation is

$$I_N[f] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i),$$

where $\{\mathbf{x}_i\}$ is a sequence of random points in H^n of length N. The approximation $I_N[f]$ converges to $I[f]$ with probability 1.

Consider an integration error ε defined as

$$\varepsilon = |I[f] - I_N[f]|.$$

The expectation of ε^2 is

$$E(\varepsilon^2) = \frac{\sigma^2(f)}{N},$$

where $\sigma^2(f)$ is the variance. The root mean square error of the MC method is

$$\varepsilon_{MC} = (E(\varepsilon^2))^{1/2} = \frac{\sigma(f)}{N^{1/2}}.$$

In contrast to grid methods, the convergence rate of MC methods does not depend on the number of variables n although it is rather slow.

The efficiency of MC methods is determined by the properties of the random numbers. Random number sampling is prone to clustering. As new points are added randomly, they do not necessarily fill the gaps between already sampled points. In contrast, low-discrepancy sequences (LDS) are specifically designed to place sample points as uniformly as possible.

The discrepancy is the measure of deviation from uniformity. Consider a number of points N from a sequence $\{\mathbf{x}_i\}$ in an n-dimensional rectangle Q whose sides are parallel to the coordinate axes, $Q \in H^n$. Then, the discrepancy is defined as

$$D_N = \sup_{Q \in H^n} \left| \frac{N_Q}{N} - m(Q) \right|,$$

where $m(Q)$ is a volume of Q and N_Q is the number of points of the sequence $\{x_i\}$ that are contained in Q.

The Koksma–Hlawka inequality gives an upper bound for the QMC integration error:

$$\varepsilon_{QMC} \leq V(f) D_N. \tag{1}$$

Here, $V(f)$ is the variation of $f(\mathbf{x})$ in the sense of Hardy and Krause [9]. For a one-dimensional function with a continuous first derivative it is simply

$$V(f) = \int_{H^1} |df(x)/dx|\, dx. \tag{2}$$

In higher dimensions, the Hardy-Krause variation may be defined in terms of the integral of partial derivatives. Further it is assumed that $f(\mathbf{x})$ is a function of bounded variation.

For random numbers, the expected discrepancy is $D_N = O((\ln\ln N)/N^{1/2})$, while the discrepancy of LDS is of the order

$$D_N = O\left(\frac{\log^n(N)}{N}\right). \tag{3}$$

There are a few well-known and commonly used LDSs. Different principles were used for their construction by Halton, Faure, Sobol, Niederreiter and others. The LDS developed by Niederreiter has the best theoretical asymptotic properties [9]. However, many practical studies have proven that the Sobol' LDS is in many aspects superior to other LDS [6,10].

The Sobol' LDS was constructed by following the three main requirements [11]:

1. Best uniformity of distribution as $N \to \infty$.
2. Good distribution for fairly small initial sets.
3. A very fast computational algorithm.

Points generated by the Sobol' LDS produce a very uniform filling of the space even for a rather small number of points N, which is a very important point in practice.

The bound on the integration error (1) is a weak one and is not particularly meaningful in practice. It was shown experimentally that the QMC integration error is determined by the variance and not by the variation of the integrand [12]. It is generally accepted that the rate of the discrepancy determines the expected rate of the accuracy, so one can use an estimate of the QMC convergence rate

$$\varepsilon_{QMC} = O\left(\frac{\log^n(N)}{N}\right). \tag{4}$$

Asymptotically, this rate of convergence is $O(N^{-1})$. Numerous computational studies showed that QMC methods can provide significant improvement over MC. The analysis of (4) shows that ε_{QMC} is an increasing function of N up to some threshold value of N^*, $N^* \approx \exp(n)$. The accelerated convergence rate $O(N^{-1})$ sets in at