Structural Safety 32 (2010) 281–292

Contents lists available at ScienceDirect

Structural Safety

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

Implementing advanced Monte Carlo simulation under spreadsheet environment

S.K. Au *, Z.J. Cao [1], Y. Wang [2]

Department of Building and Construction, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Available online 8 April 2010

Keywords:
Markov Chain Monte Carlo
Monte Carlo
Reliability method
Spreadsheet
Subset Simulation
VBA

ABSTRACT

This paper presents a spreadsheet computational framework for implementing an advanced Monte Carlo method called Subset Simulation for uncertainty propagation that can provide better resolution for low failure probability level at the same time retaining some robustness features of direct Monte Carlo. While the efficiency of Subset Simulation has been demonstrated by numerous studies, attention in this work is devoted to application robustness of the spreadsheet framework. This concern is relevant because advanced Monte Carlo algorithms, or in general variance reduction techniques, gain their efficiency by exploiting information about the problem, which may require intrusive exchange of information with the system analysis model during the simulation process. To explore and authenticate implementation issues, a prototype Visual Basic Application (VBA) package is developed that can perform efficient uncertainty propagation by plugging as an Add-In into a spreadsheet that performs deterministic analysis. The resulting uncertainty propagation process is non-intrusive, requiring immaterial modification of the deterministic analysis spreadsheet. Operationally the proposed framework divides the whole process into system modeling (deterministic analysis), uncertainty modeling (generation of random variables) and uncertainty propagation (Subset Simulation). It is hoped that the development work can promote the use of advanced Monte Carlo simulation tools for uncertainty propagation in the decision-making process.

1. Introduction

Uncertainty propagation in engineering systems involves the probabilistic assessment of system performance attributes given a probabilistic modeling of the significant uncertainties and proper deterministic modeling of system behavior [1–6]. It involves calculating the statistical average of response quantities of interest which are formulated as probability integrals over the space of random parameters. From a risk assessment perspective, the distribution tail of performance attributes is often of interest. The performance of a proposed design in the presence of uncertainty can be quantified in terms of the 'failure probability' or 'performance margin' with respect to specified design objectives. Let $\underline{\Theta} \in R^n$ denote the vector of random variables for which a probability model is available, say, in terms of the joint probability density function (PDF) $p(\underline{\theta})$. Without loss of generality, we assume that the random variables are independent. After all, in applications correlated random variables are generated by independent ones through transformations. Many failure events in engineering risk analysis can be formulated as the exceedance of a critical response variable $Y(\underline{\Theta})$ over some specified threshold level y [7], i.e.,

$$P(F) = P(Y > y) = \int_F p(\underline{\theta}) d\underline{\theta} \tag{1}$$

Complementary to the failure probability is the performance margin that corresponds to the percentile of a given risk tolerance through which the risk tolerance of a decision maker manifests. For example, the 90-, 99-, and 99.9-percentiles may provide a decision maker with low-, medium-, and high-confidence estimates in the upper-bound value of response, corresponding to a risk-tolerant, risk-neutral, or risk-averse decision maker, respectively.

In realistic projects the system models to be analyzed have been evolving with higher complexity and wider variety. Insights about system behavior become less available and more costly to digest, making analytical or semi-analytical solutions almost intractable. In this context a robust uncertainty propagation method whose applicability is insensitive to the problem nature and complexity is most desirable. Direct Monte Carlo [8] is the most robust method but it suffers from a lack of resolution at the distribution tail, i.e., rare event, which is often of interest in risk assessment problems. Advanced reliability methods, often called 'variance reduction techniques', have been developed over the years [9–13]. Their general goal has been to minimize the number of system analyses (i.e., computation of system response) for producing estimates of

* Corresponding author. Tel.: +852 2194 2769; fax: +852 2788 7612.
E-mail addresses: siukuiau@cityu.edu.hk (S.K. Au), zijuncao3@student.cityu.edu.hk (Z.J. Cao), yuwang@cityu.edu.hk (Y. Wang).
[1] Tel.: +852 3442 6492.
[2] Tel.: +852 2788 7605; fax: +852 2788 7612.

0167-4730/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.strusafe.2010.03.004

acceptable statistical accuracy. There remains no general efficient algorithm available, however, especially for complex systems. The main difficulties come from: (1) the relationship between the random variables and the performance attributes is only implicitly known through point-wise system analysis; (2) there are many random variables involved in the system; (3) information about rare scenarios is required; and (4) there are many interactive response variables in the description of performance criteria. The first factor renders the system a 'black-box', requiring higher computational efforts to understand its behavior. The second one increases the dimension of integration of any statistical quantity of interest. Any computational procedure based on numerically integrating out the individual dimensions requires exponentially growing computational effort and hence is not efficient for dimensions greater than three.

Models for analyzing complex systems are characterized by a large number of degrees of freedom, time-varying and response-dependent nonlinear behaviour and increasingly governed by multi-disciplinary laws. Although the advent of computer technology has allowed the analysis of complex systems for a given scenario (i.e., one sample) to be performed with affordable computational time, the same is not true for uncertainty propagation, since the latter involves repeated analysis. Even if resources are available, they should be deployed in a cost-effective manner that yields most information on failure scenario of concern. In general, there is a trade-off between efficiency and robustness of an advanced simulation algorithm. This has motivated the recent development of efficient yet robust algorithms for propagating uncertainties in complex systems [12].

Software implementation of structural reliability algorithms has received greater attention in parallel with development of reliability methods. Software packages have been developed by different institutions with different objectives and target applications, e.g., COSSAN [14], STRUCTUEL [15], CALREL [16] and NESSUS [17]. Excellent review can be found in a recent special issue [18]. Motivated by the work of Low and Tang [19–21] that developed spreadsheet implementation of first order reliability method in structural and geotechnical engineering, this paper explores the spreadsheet implementation of an advanced Monte Carlo method called Subset Simulation [22,23]. The motivation for spreadsheet implementation stems from the wide variety of practical problems and models that can be analyzed using spreadsheet in modern office settings, spanning applications in geotechnical engineering [4], system engineering [24], financial engineering [25], etc. On the other hand, while the spreadsheet implementation of direct Monte Carlo has been well demonstrated in commercial codes, e.g., @Risk [26], it is the intent of this work to explore the implementation of Monte Carlo algorithms of an advanced nature. The basic concern lies in whether, or to what extent, one can decouple the modeling calculations from those required by the advanced simulation algorithm, since it is commonly recognized that the latter gains efficiency over direct Monte Carlo by exploiting specific knowledge about the system that often requires intrusive (and possibly problem-dependent) exchange of information during simulation. It shall be demonstrated that in the case of Subset Simulation it is possible to decouple the spreadsheet development into uncertainty modeling (generation of random variables), system modeling (deterministic analysis), and uncertainty propagation (Subset Simulation), thereby allowing uncertainty propagation to be performed in a non-intrusive robust manner.

The basic theory of Subset Simulation will be presented first, followed by the framework of spreadsheet implementation. A slope stability risk example in geotechnical engineering will be used to illustrate the application of the proposed framework via an in-house developed EXCEL Add-In.

2. Subset Simulation

Subset Simulation is an adaptive stochastic simulation procedure for efficiently computing small tail probabilities [22,23]. Originally developed for dynamic reliability analysis of building structures, it stems from the idea that a small failure probability can be expressed as a product of larger conditional failure probabilities for some intermediate failure events, thereby converting a rare event simulation problem into a sequence of more frequent ones. During simulation, conditional samples are generated from specially-designed Markov chains so that they populate gradually each intermediate failure region until they reach the final target (rare) failure region.

Let Y be a given critical response for which $P(Y > y)$ is of interest, and $0 < y_1 < y_2 < \cdots < y_m = y$ be an increasing sequence of intermediate threshold values. It should be noted that considering a single critical response leads to little loss of generality because multiple failure criteria can be incorporated into a single one [22]. By sequentially conditioning on the event $\{Y > y_i\}$, the failure probability can be written as.

$$P(Y > y) = P(Y > y_1) \prod_{i=2}^{m} P(Y > y_i | Y > y_{i-1}) \tag{2}$$

The raw idea stemming from this expression is to estimate $P(Y > y_1)$ and $\{P(Y > y_i | Y > y_{i-1})\colon i = 2, \cdots, m\}$ by generating samples of $\underline{\Theta}$ conditional on $\{Y(\underline{\Theta}) > y_i\colon i = 1, \cdots, m\}$. In the actual implementation, $y_1, \cdots, y_m$ are generated adaptively using information from simulated samples so that the sample estimate of $P(Y > y_1)$ and $\{P(Y > y_i | Y > y_{i-1})\colon i = 2, \cdots, m\}$ always correspond to a common specified value of the conditional probability p_0 (p_0=0.1 is found to be a good choice). Subset Simulation is in fact a procedure for generating estimates of performance margins corresponding to pre-specified failure probability levels.

The efficient generation of conditional samples is highly-non-trivial but pivotal to the success of Subset Simulation. It is made possible through the machinery of a class of powerful algorithms called Markov Chain Monte Carlo (MCMC) simulation [27–29]. In MCMC, successive samples are generated from a specially-designed Markov chain whose limiting stationary distribution tends to the target PDF as the length of the Markov chain increases. An essential aspect of the implementation of MCMC is the choice of 'proposal distribution' that governs the generation of the next sample from the current one and consequently the efficiency of the algorithm. We shall come back to this after describing the overall simulation procedure.

The Subset Simulation procedure for adaptively generating samples of $\underline{\Theta}$ conditional on $\{Y(\underline{\Theta}) > y_i\colon i = 1, \cdots, m\}$ corresponding to specified target probabilities $\{P(Y(\underline{\Theta}) > y_i) = p_0^i\colon i = 1, \cdots, m\}$ is summarized as follows. First, N samples $\{\underline{\Theta}_{0,k}\colon k = 1, \cdots, N\}$ are simulated by direct Monte Carlo, i.e., they are i.i.d. as the original PDF. The subscript '0' here denotes that the samples correspond to 'conditional level 0' (i.e., unconditional). The corresponding values of the response $\{Y_{0,k}\colon k = 1, \cdots, N\}$ are then computed. The value of y_1 is chosen as the $(1 - p_0) \cdot N$th value in the ascending list of $\{Y_{0,k}\colon k = 1, \cdots, N\}$, so that the sample estimate for $P(F_1) = P(Y > y_1)$ is always equal to p_0. Here, it is assumed that p_0 and N are chosen such that $p_0 \cdot N$ is an integer.

Due to the choice of y_1, there are $p_0 \cdot N$ samples among $\{\underline{\Theta}_{0,k}\colon k = 1, \cdots, N\}$ whose response Y lies in $F_1 = \{Y > y_1\}$. These are samples at 'conditional level 1' and are conditional on F_1. Starting from each of these samples, MCMC is used to simulate additional $(1 - p_0) \cdot N$ conditional samples so that there are a total of N conditional samples at conditional level 1. The value of y_2 is then chosen as the $(1 - p_0) \cdot N$th value in the ascending list of $\{Y_{1,k}\colon k = 1, \cdots, N\}$, and it defines $F_2 = \{Y > y_2\}$. Note that the sample