

Correlated Random Variables in Probabilistic Simulation

Miroslav Vořechovský, MSc.

Drahomír Novák, Assoc. Prof. DrSc.

Summary

A new efficient technique to impose the statistical correlation when using Monte Carlo type method for statistical analysis of computational problems is proposed. The technique is based on stochastic optimization method called Simulated Annealing. The comparison with other techniques presently used and intensive numerical testing showed the superiority and robustness of the method. No significant obstacles have been found working also with large problems (large number of random variables). Advantages and limitations of the approach will be discussed. Remarks on positive definiteness of target correlation matrix will be made. Numerical examples show the efficiency of the method.

Keywords: Statistical correlation, Monte Carlo, LHS, simulated annealing.

1. Introduction

The aim of statistical and reliability analysis of any computational problem which can be numerically simulated is mainly the estimation of statistical parameters of response variable and/or theoretical failure probability. Pure Monte Carlo simulation (MC) cannot be applied for time-consuming problems, as it requires large number of simulations (repetitive calculation of response). Small number of simulations can be used for acceptable accuracy of statistical characteristics of response using stratified sampling technique Latin Hypercube Sampling (LHS) [1, 2, 3, 4, 5, 6]. LHS strategy has been used by many authors in different fields of engineering. The classical reliability theory introduced the basic concept formally using the response variable $Z = g(\mathbf{X})$, where g (computational model) represents functional relationship between elements of vector \mathbf{X} (uncertainties - random variables). The paper is focused on the problem of efficient imposition of statistical correlation into quantities of the vector \mathbf{X} within framework of MC (preferably LHS). Techniques presently available are discussed first.

2. Sampling and Statistical Correlation

The N_{Sim} samples (where N_{Sim} is number of simulations planned for each random variable X_i) are chosen from cumulative distribution function (CDF) domain in different ways, e. g. randomly by inverse transformation of CDF. In following we presume using LHS methodology for sampling. Table 1 represents sampling scheme, where simulation numbers are in rows and columns are related to random variables and N_V is number of input variables.

Table 1 Sampling scheme for N_{Sim} deterministic calculations of $\mathbf{g}(\mathbf{X})$

Simulation	Var. 1	Var. 2	Var. 3	...	Var. N_V
1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$...	$x_{1,NV}$
2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$...	$x_{2,NV}$
...	$x_{3,NV}$
N_{Sim}	$x_{NSim,1}$	$x_{NSim,2}$	$x_{NSim,3}$...	$x_{NSim,NV}$

There are generally two problems related to statistical correlation: First, during sampling undesired correlation can be introduced between random variables (columns in Table 1). For example instead a correlation coefficient zero for uncorrelated random variables undesired correlation, e.g. 0.6 can be generated. It can happen especially in case of very small number of simulations (tens), where the number of combinations is rather limited. Second task is to introduce prescribed statistical correlation between random variables defined by correlation matrix. Columns in Table 1 should be rearranged in such way to fulfill these two requirements: to diminish undesired random correlation and to introduce prescribed correlation. The efficiency of LHS technique was showed first time in [1], but only for uncorrelated random variables. A first technique for generation of correlated random variables has been proposed by [4]. The authors showed also the alternative to diminish undesired random correlation. The technique is based on iterative updating of sampling matrix; Cholesky decomposition of correlation matrix has to be applied. The technique can result in a very low correlation coefficient (absolute value) if generating uncorrelated random variables. But authors of [5] have found that the approach tends to converge to an ordering which still gives significant correlation errors between some variables. The scheme has more difficulties when simulating correlated variables. The correlation procedure can be performed only once, there is no way to iterate it and to improve the result. These obstacles stimulated the authors of [5], they proposed so called single-switch-optimized ordering scheme. Their approach is based on iterative switching of the pair of samples of Table 1. The authors showed that theirs technique clearly performs well

enough, but it may still converge to a non-optimum ordering. A different method is needed for simulation of both uncorrelated and correlated random variables. Such method should be efficient enough: reliable, robust and fast.

Note that the accurate best result is obtained if all possible combinations of ranks for each column (variable) itself in Table 1 are treated. It would be necessary to try extremely large number of rank combinations $(N_{Sim})^{N^2}$. It is clear that this rough approach is hardly applicable in spite of the fast development of computer hardware. Note that we leave the concept of samples selection from N -dimensional marginal PDF (with different partial components) and prescribed covariance structure (correlation matrix).

3. Stochastic Optimization *Simulated Annealing*

The imposition of prescribed correlation matrix into sampling scheme can be understood as an optimization problem: The difference between prescribed \mathbf{K} and generated \mathbf{S} correlation matrices should be as small as possible. A suitable measure of quality of overall statistical properties can be introduced, e.g. the maximal difference of correlation coefficients between matrices E_{max} or a norm, which takes into account deviations of all correlation coefficients:

$$E_{max} = \max_{1 \leq i < j \leq N_V} |S_{i,j} - K_{i,j}|, \quad E_{overall} = \sqrt{\sum_{i=1}^{N_V-1} \sum_{j=i+1}^{N_V} (S_{i,j} - K_{i,j})^2} \quad (1)$$

The norm E has to be minimized from the point of view of definition of optimization problem: the objective function is E and the design variables are related to ordering in sampling scheme (Table 1). It is well known that deterministic optimization techniques and simple stochastic optimization approaches can very often fail to find the global minimum. Such technique fails in some local minimum and then there is no chance to escape from it and to find the global minimum. It can be intuitively predicted that in our problem we are definitely facing the problem with multiple local minima. Therefore we need to use the stochastic optimization method, which works with some probability of escaping from local minimum. The simplest form is the two-member evolution strategy, which works in two steps: Mutation and selection.

1. Step 1 (mutation): In generation a new arrangement of random permutations matrix \mathbf{X} is obtained using random changes of ranks, one change is applied for one random variable. Generation should be performed randomly. Objective function (norm E) can be then calculated using newly

obtained correlation matrix - it is usually called “offspring”. The norm E calculated using former arrangement is called “parent”.

2. Step 2 (selection): The selection chooses the best norm between the “parent” and “offspring” to survive: For the new generation (permutation table arrangement) the best individual (table arrangement) has to give a value of objective function (norm E) smaller than before.

Such approach has been intensively tested using numbers of examples. It was observed that the method in most cases could not capture the global minimum. It failed in a local minimum and there was no chance to escape from it, as only improvement of the norm E resulted in acceptance of “offspring”.

The step “Selection” can be improved by Simulated Annealing approach (SA), a technique, which is very robust concerning the starting point (initial arrangement of permutations table). The SA is optimization algorithm based on randomization techniques and incorporates aspects of iterative improvement algorithms. Basically it is based on the Boltzmann probability distribution:

$$P_r(E) \approx e^{\left(\frac{-\Delta E}{k_b \cdot T}\right)} \quad (2)$$

where ΔE is difference of norms E before and after random change. This probability distribution expresses the concept when a system in thermal equilibrium at temperature T has its energy probabilistically distributed among all different energy states ΔE . Boltzmann constant k_b relates temperature and energy of the system. Even at low temperatures, there is a chance (although very small) of a system being locally in a high energy state. Therefore, there is a corresponding possibility for the system to move from a local energy minimum in favor of finding a better minimum (escape from local minimum). There are two alternatives in step 2 (mutation). *Firstly*, new arrangement (offspring) results in decrease of the norm E . Naturally offspring is accepted for new generation. *Secondly*, offspring does not decrease the norm E . Such offspring is accepted with some probability (2). This probability changes as temperature changes. As the result, there is a much higher probability that the global minimum is found in comparison with deterministic methods and simple evolution strategies.

In our case k_b can be considered to be one. In classical application of SA approach for optimization there is one problem: how to set the initial temperature? Usually it should be considered heuristically. But our problem is constrained: Acceptable elements of correlation matrix are always from interval $\langle -1; 1 \rangle$. Based on this fact the maximum of the norm (1) can be

estimated using prescribed and hypothetically “most remote” (unit correlation coefficients, plus or minus). This approach represents a significant advantage: The heuristic estimation of initial temperature is neglected; the estimation can be performed without the guess of the user and the “trial and error” procedure. The initial temperature has to be decreased step by step, e.g. using reduction factor f_T after constant number of iterations (e.g. thousand) $T_{i+1} = T_i * f_T$. The simple case is to use e.g. $f_T = 0.95$, note that more sophisticated cooling schedules are known in SA-theory [7, 8].

4. Numerical Examples

4.1 Correlated Properties of Concrete

In order to illustrate the efficiency of proposed technique, consider an example of correlation matrix, which corresponds to properties of a concrete. They are described by 7 random variables; parametric study of this example is given with emphasis on influence of number of simulations is given in [6]. It can be seen [6] that as number of simulations increases, correlation matrix is closer to the target one. Using standard PC correlating with SA took about one second. Fig. 1 shows the decrease of norm E during SA-process. Such figure is typical and should be monitored.

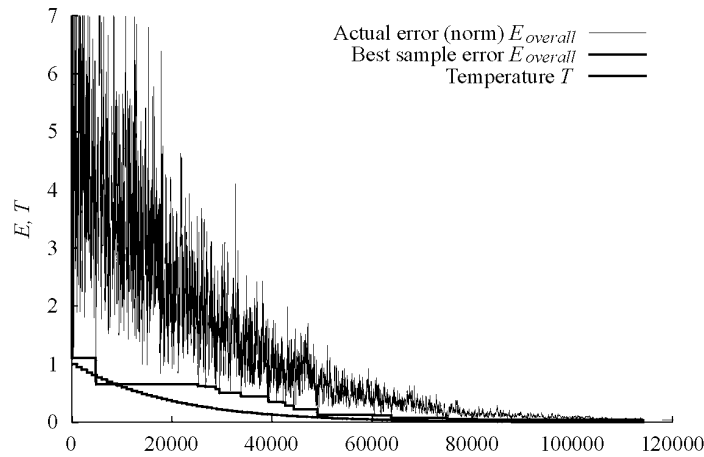


Fig. 1 The norm E (error) vs. number of random changes (rank switches).

Another example of utilization is given in [9] (this proceedings), where generation of uncorrelated random variables was needed to represent material

strength over a structure instead of random field approach.

In real applications of simulation technique in engineering (e.g. LHS), statistical correlation represents very often a weak part of a priori assumptions. Because of this pure knowledge a prescribed correlation matrix on input can be non-positive definite. The user can have difficulties to update correlation coefficients in order to make the matrix positive definite. The example presented here demonstrates that when non-positive definite matrix is on input, SA can work with it and resulting correlation matrix is as close as possible to originally prescribed matrix but the dominant constraint (positive definiteness) is satisfied automatically. Consider a very unrealistic simple case of statistical correlation for three random variables A, B a C according to the matrix \mathbf{K} (columns and rows correspond to the ranks of variables A, B, C):

$$\mathbf{K} = \begin{pmatrix} \mathbf{1} & 0.9 & 0.9 \\ & \mathbf{1} & -0.9 \\ \text{symm} & & \mathbf{1} \end{pmatrix}, \quad \mathbf{S}_1 = \begin{pmatrix} \mathbf{1} & 0.499 & 0.499 \\ & \mathbf{1} & -0.499 \\ \text{symm} & & \mathbf{1} \end{pmatrix} \begin{bmatrix} 0.401 \\ \mathbf{0.695} \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{pmatrix} \mathbf{1} & 0.311 & 0.311 \\ 1 & \mathbf{1} & -0.806 \\ 1 & \underline{10} & \mathbf{1} \end{pmatrix} \begin{bmatrix} 0.589 \\ \mathbf{0.884} \end{bmatrix}, \quad \mathbf{S}_3 = \begin{pmatrix} \mathbf{1} & 0.236 & 0.236 \\ 1 & \mathbf{1} & -0.888 \\ 1 & \underline{100} & \mathbf{1} \end{pmatrix} \begin{bmatrix} 0.644 \\ \mathbf{0.947} \end{bmatrix}$$

The correlation matrix is obviously not positive definite. Strong positive statistical correlation is required between variables (A, B) and variables (A, C), but strong negative correlation between variables (B, C). It is clear that only compromise solution can be done. The method resulted in such compromise solution without any problem, \mathbf{S}_1 (number of simulations N_{Sim} was high enough to avoid limitation in number of rank combinations). Final values of norms are included on the right side: first line corresponds to norm E_{max} (1) second line (bold) means overall norm $E_{overall}$ (1). This feature of the method can be accepted and interpreted as an advantage of the method. In practical reliability problems with non-positive definiteness exist (lack of knowledge). It represents the limitation when using some other methods (Cholesky decomposition of prescribed correlation matrix).

In real applications it can be a greater confidence to one correlation coefficient (good data) and a smaller confidence to another one (just estimation). Solution of mentioned problems is weighted calculations of both norms (1). For example the norm $E_{overall}$ (1) can be modified in this way:

$$E_{overall} = \sqrt{\sum_{i=1}^{N_V-1} \sum_{j=i+1}^{N_V} w_{i,j} \cdot (S_{i,j} - K_{i,j})^2} \quad (3)$$

where $w_{i,j}$ is weight of certain correlation coefficient. Several examples of choices and resulting correlation matrices (with both norms) are above. Resulting matrices S_2 and S_3 illustrate significance of proportions among weights. Weights are in lower triangle and matrix K is targeted again. Weights of accentuated members and resulting values are underlined.

5. Conclusions

The new efficient technique of imposing the statistical correlation when using Monte Carlo type simulation is suggested. The technique is robust, efficient and very fast. The method is implemented in a multipurpose software package FREET based on LHS for reliability analysis of computational problems. The method has several advantages in comparison with former techniques:

1. The technique uses only random changes of ranks in sampling matrix. Number of simulations does not increase CPU time in practical cases, but for increasing number of random variables more SA simulations is needed to achieve a good accuracy. The technique is robust, Simulated Annealing can be terminated if the error (norm) is acceptable (users decision).
2. The problem of imposing statistical correlation is constrained precisely; therefore the parameters for annealing can be estimated.
3. The technique enables emphasizing of important coefficients using weights while others can be suppressed.

Acknowledgement

The authors thank for support under the grant of Grant Agency of the Czech Republic No. 103/02/1030 and CEZ J22/98:261100007.

References

- [1] MCKAY M.D., CONOVER W.J. & BECKMAN R.J.: "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21, 1979, 239-245
- [2] IMAN R.C. & CONOVER W.J.: "Small Sample Sensitivity Analysis Techniques for Computer Models with an Application to Risk Assessment". *Communications in Statistics: Theory and Methods*, A9, 17, 1980, 1749-1842
- [3] NOVÁK D., TEPLÝ B. & KERŠNER Z.: "The role of Latin Hypercube Sampling method in reliability engineering". *Proceedings of ICOSSAR-*

- 97, Kyoto - Japan, Editors Shiraishi N., Shinozuka M., Wen Y.K., Rotterdam: Balkema, 1998. 403-409
- [4] IMAN R.C. & CONOVER W.J.: "A Distribution Free Approach to Inducing Rank Correlation Among Input Variables". *Communications in Statistics. B11*, 1982, 311-334
- [5] HUNTINGTON D.E. & LYRINTZIS C.S.: "Improvements to and limitations of Latin hypercube sampling". *Probabilistic Engineering Mechanics 13,4*, 1998, 245-253
- [6] VOŘECHOVSKÝ M., NOVÁK D., RUSINA R.: "A New Efficient Technique for Samples Correlation in Latin Hypercube Sampling", *Proc. of 7th International Scientific Conference*, Košice, Slovakia, 2002, 102-108
- [7] OTTEN R. H. J. M. & GINNEKEN L. P. P. P: *The Annealing Algorithm*. Kluwer Academic Publishers, USA 1989
- [8] LAARHOVEN P.J & AARTS E.H.: *Simulated Annealing: Theory and Applications*. D. Reidel Publishing Company, Holland, 1987
- [9] LEHKÝ D., NOVÁK D.: Nonlinear Fracture Mechanics Modeling of Size Effect in Concrete under Uniaxial Tension. *Proceedings of 4th International PhD Symposium in Civil engineering*. Munich - Germany 2002. (In print)



Miroslav Vořechovský, MSc.
PhD Candidate

University of Technology,
Faculty of Civil Engineering,
Institute of Structural Mechanics
Veveří 95, 662 37
Czech Republic

Tel.: +420 5 41 14 71 31
Fax: +420 5 41 24 09 94



Assoc. Prof. Drahomír Novák, DrSc.
Supervisor

University of Technology
Faculty of Civil Engineering
Institute of Structural Mechanics
Veveří 95, 662 37
Czech Republic