On Correlation Control in Monte Carlo type Sampling

Miroslav Vořechovský
Institute of Structural Mechanics, Faculty of Civil Engineering, Brno University of Technology, Veveří 95, 602 00 Brno, Czech Republic, vorechovsky.m@fce.vutbr.cz

Abstract: The objective of this paper is a study of performance of correlation control of recently proposed procedure for sampling from a multivariate population within the framework of Monte Carlo simulations (especially Latin Hypercube Sampling). In particular, we study the ability of the method to fulfill the prescribed correlation structure of a random vector for various sample sizes and number of marginal variables. Two norms of correlation error are defined, one very conservative and related to extreme errors and other related to averages of correlation errors. We study the behavior of Pearson correlation coefficient for Gaussian vectors and Spearman rank order coefficient (as a distribution-free correlation measure). Theoretical results on performance bounds for both correlation types in the case of desired uncorrelatedness are compared to performance of the proposed technique and also to other previously developed techniques for correlation control, namely the Cholesky orthogonalization as applied by Iman and Conover (1980,1982); and Gram-Schmidt orthogonalization used by Owen (1994).

Keywords: statistical correlation distribution, uncorrelatedness, orthogonality, orthogonal vectors, multivariate Monte Carlo simulation, Latin Hypercube Sampling, correlation performance, saturated design

1. Introduction

The aim of statistical and reliability analyses of any computational problem which can be numerically simulated is mainly the estimation of statistical parameters of response variables and/or theoretical failure probability. Pure Monte Carlo simulation cannot be applied to time-consuming problems as it requires a large number of simulations (repetitive calculation of responses). A small number of simulations can be used to gain an acceptable level of accuracy for the statistical characteristics of the response using the stratified sampling technique Latin Hypercube Sampling (LHS) first developed by Conover (1975) and later elaborated mainly by McKay at al. (1979) and Iman and Conover (1980).

It is known that the output response variables of some systems, represented by their response functions, are sensitive to changes in correlations among the input variables. Therefore, it is essential to precisely capture the input correlations in the simulated values. Thus, Monte Carlo type simulation approaches require sampling of correlated data from Gaussian and frequently also non-Gaussian distributions. Other than the multivariate normal distribution, few random-vector models are tractable and general, though many multivariate distributions are well documented (Johnson 1987).

In the present paper, the task of correlation control in sampling is treated as a combinatorial optimization problem. The technique was developed in (Vořechovský 2002) and (Vořechovský and Novák 2002, 2003), and since then it was improved in (Vořechovský 2007, Vořechovský and Novák 2009). In the technique, an analogy between the statistical mechanics of large multivariate physical systems and combinatorial optimization is used to develop a strategy for the optimal ordering of samples to control the

4th International Workshop on Reliable Engineering Computing (REC 2010)
Edited by Michael Beer, Rafi L. Muhanna and Robert L. Mullen
Copyright © 2010 Professional Activities Centre, National University of Singapore.
ISBN: 978-981-08-5118-7. *Published by* Research Publishing Services.
doi:10.3850/978-981-08-5118-7_023

correlation structure. The problem of optimal sample ordering is solved by the so-called Simulated Annealing method using a Monte Carlo procedure similar to the one developed by Metropolis et al. (1953).

The technique is designed to minimize differences between the desired (target) correlation matrix and actual (estimated) correlation matrix in samples generated by any Monte Carlo type method (e.g. Latin Hypercube Sampling – LHS). In this paper, performance studies for correlation control are presented and the performance is compared to theoretical results obtained earlier by the author. In this way, the present paper promotes the results obtained in (Vořechovský 2006, 2007, 2009a, 2009b). As mentioned before, the technique works with arbitrarily sampled data. However, in the remainder of the paper, we assume that samples are generated using Latin Hypercube Sampling; in particular its special alternative called LHS-mean by Vořechovský and Novák (2009).

2. Combinatorial optimization for correlation control

2.1. SAMPLING CORRELATION (CORRELATION ESTIMATION)

The sampling correlation is assumed to be estimated by one of the standard techniques among which the most spread ones are the linear Pearson correlation coefficient, Spearman rank-order correlation and Kendall's tau. A detailed information on the computation of these correlations can be found e.g. in (Vořechovský 2007, 2009a, 2009b). The sample is represented by a table of dimensions N_{var} times N_{sim} similar to the one at the right hand side of Fig. 1. Each row corresponds to one variable each of which is represented by a sample (row vector) of size N_{sim}. The related square and symmetric correlation matrices have therefore the order of N_{var}.

The estimated correlation (matrix **A**) of all pairs of N_{var} variables is computed using the sample of size N_{sim}. This actual correlation is compared with the target correlation matrix **T** that is supposed to somehow describe the desired dependency pattern.

2.2. CORRELATION ERROR MEASURES

The imposition of the prescribed correlation matrix into a sampling scheme can be understood as an optimization problem: we want to *minimize the difference* between the target correlation matrix (e.g. user defined, prescribed) **T** and the actual correlation matrix **A**. Let us denote the difference matrix (error-matrix) **E**:

$$\mathbf{E} = \mathbf{T} - \mathbf{A} \tag{1}$$

To have a scalar measure of the error we introduce a suitable norm of the matrix **E**. In particular, a good and conservative measure of the distance between **T** and **A** can be the norm defined as:

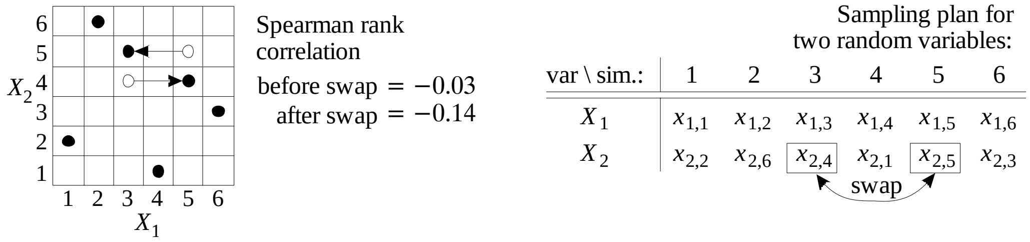
$$\rho_{\max} = \max_{1 \le i < j \le N_{\text{var}}} \left| E_{i,j} \right| \tag{2}$$

Even though this norm has a clear meaning and known "units" (correlation) and can be used as a good stopping condition in the iterative algorithm, it is not a suitable objective function to be subjected to direct minimization. The reason is that it represents the maximum distance from the origin along any correlation difference in the $N_c = N_{var}(N_{var} - 1)/2$ dimensional space of all correlations represented by the error matrix **E**, see Vořechovský (2009a). A better choice is a norm taking into the account deviations of all correlation coefficients:

$$\rho_{\text{rms}} = \sqrt{\frac{2}{N_{\text{var}}\left(N_{\text{var}}-1\right)}}\sqrt{\sum_{i=1}^{N_v-1}\sum_{j=1}^{N_{\text{var}}}\left(E_{i,j}\right)^2} \tag{3}$$

where we used the symmetry of the correlation matrices by summing up the squares of the upper triangle off-diagonal terms only. This norm measures the *distance* of the actual correlation error (a point) from the origin in the space of all N_c different correlations, i.e. the hypotenuse multiplied by the first square root appearing in Eq. (3). This norm proved itself to be a good objective function for the optimization algorithm described below. Taking the square root yields a measure in units of correlation which represents the normalized error per entry and is, therefore, suitable for comparison when examples of a different number of variables N_{var} are involved.

The norm ρ_{rms} has to be minimized, from the point of view of definition of the optimization problem; the *objective function* is ρ_{rms} and the *design variables* are related to the *ordering* in the sampling scheme (Fig. 1). Clearly, in real applications the space of the possible actual correlation matrices **A** is extremely large: consider all $\left(N_{\text{sim}}!\right)^{N_{\text{var}}-1}$ different mutual orderings of the sampling table. Clearly, we want to find an *efficient near-optimal solution*. This is believed to be achieved by application of the algorithm briefly described next.



var \ sim.:	1	2	3	4	5	6
X_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$
X_2	$x_{2,2}$	$x_{2,6}$	$\boxed{x_{2,4}}$	$x_{2,1}$	$\boxed{x_{2,5}}$	$x_{2,3}$

Figure 1. Illustration of a random trial – swap of samples *j* and *k* of variable X_2.

2.3. OPTIMIZATION ALGORITHM FOR RANKS SHUFFLING

In each step of the combinatorial optimization algorithm, mutation is performed by a transition called a *swap* from the parent configuration to the offspring configuration. A swap (or a *trial*) is a small change to the arrangement of sampling table in Fig. 1. It is done by randomly interchanging a pair of two values $x_{i,j}$ and $x_{i,k}$. In other words one needs to randomly generate *i* (select the variable), and a pair *j*,*k* (select the pair of realizations to interchange), see Fig. 1. One swap may or may not lead to a decrease (improvement) in the norm. Immediately, one configuration between the parent and offspring is *selected* to survive. Simulated Annealing algorithm is employed for the selection step. The advantage of this compared to some simple evolution strategy is that there is a nonzero probability of accepting an offspring configuration with a higher error than its parent (hill climbing). The acceptance rule with decaying probability of hill climbing gives us a mechanism for accepting increases in a controlled fashion (cooling schedule). It is possible that accepting an increase in the penalty function will reveal a new configuration that will avoid a local minimum or at least a bad local minimum in future. Details on the algorithm an also details on the implementation can be found in (Vořechovský 2002, 2007) and (Vořechovský and Novák 2002, 2003, 2009).

3. Theoretical bounds on correlation errors

3.1. CORRELATION ERRORS FOR A RANDOM ORDERING

Obviously, the upper bound on any of the correlation measure must be one (an extreme absolute value of correlation). Let us tighten the bound. We, however, can say that, on average, the upper bound on correlation error is the one that occurs when *randomly shuffling* the ranks without any control.

The described algorithm starts with a random permutation and at the end of the process selects the best available solution so far. Therefore, the upper bound of an average error is the one related to a random permutation. It is a well known fact, that in the case of a pair of random variables, both Pearson and Spearman correlation coefficient are, in the limit, normally distributed with zero mean and standard deviation of $(N_{\text{sim}}-1)^{-1/2}$, a formula obtained by Student and incorporated in Pearson's (1907) memoir, see (Hotelling and Pabst 1936).

Vořechovský (2009a) has shown that the error ρ_{rms} for random ordering and in multidimensional setting follows (in the limit of $N_{\text{sim}} \to \infty$) Chi distribution with the mean value of

$$\mu_{\text{rms}} = \frac{r_\mu\left(N_{\text{var}}\right)}{\sqrt{N_{\text{sim}}-1}} \tag{4}$$

where the function $r_\mu(N_{\text{var}})$ quickly converges to unity with growing N_{var} and thus the mean error is almost independent of the dimension N_{var}. The limiting rates of the average error and its standard deviation are:

$$\underset{N_{\text{sim}}\to\infty}{\mu_{\text{rms}}} = N_{\text{sim}}^{-1/2} \quad \text{and} \quad \underset{N_{\text{sim}}\to\infty}{\sigma_{\text{rms}}} = N_{\text{sim}}^{-1/2}\ N_{\text{var}}^{-1} \tag{5}$$

Details on the distribution of ρ_{rms} can be found in (Vořechovský 2009a).

Regarding the distribution of the second norm, ρ_{max}, we deal with extremes of (approximately independent and) uniformly distributed Gaussian random variables. It has been shown (Vořechovský 2009a) that both, the mean value and standard deviation of the random error ρ_{max} (Gumbel distributed) are asymptotically proportional to $N_{\text{sim}}^{-1/2}$. The decrease of the standard deviation of a random correlation is, however, slower than in the case of ρ_{rms}. Also, the average error slowly grows with an increasing dimension N_{var}. We can say that ρ_{max} is more conservative error measure compared to ρ_{rms}. The derivation of the error distribution and formulas for the error statistics can be found in (Vořechovský 2009a).

We have shown the intuitively acceptable result that zero uncorrelatedness is the most frequent pattern among all possible random orderings. In many applications the sample size is not restricted and the analysis can be performed with extremely large N_{sim}. In such situations, if independence among input variables is requested, the analysis can be performed with a sample on which no special algorithm for dependence/correlation control is applied.

To conclude, we are sure that the proposed combinatorial algorithm performs such that any of the errors of the decays (on average) at the rate of $N_{\text{sim}}^{-1/2}$. Such an error might be satisfactory for very large sample sizes and therefore the correlation could be left random when requesting uncorrelatedness. However, in the case of a *small sample size*, one should use a suitable method for correlation control to speed up the convergence towards the zero error. Section 4 will present how efficient the proposed algorithm is.

3.2. LOWER BOUNDS ON CORRELATION ERRORS

The best solution that can be achieved is when the actual correlation matrix matches the target one, i.e. both correlation norms are zero. This, however, cannot always be achieved.

When uncorrelatedness between two random variables is requested, it can be shown (Vořechovský 2009a) by a simple pairing argument that in the case of Spearman correlation, the smallest error is $6N_{\text{sim}}^{-3}$ when the sample size is N_{sim} = 2+4l, where l is a nonnegative integer (the exact minimal error is $6N_{\text{sim}}^{-1}/\left(N_{\text{sim}}^{2}-1\right)$ and comes from the analysis of simplified formula for sample Spearman correlation). In other words, the worst convergence (of the peaks) of the smallest correlation error is polynomial (a power law with exponent of −3), and the associated error graph in a double logarithmic plot is a decreasing straight line of the same slope. For other sample sizes, the correlation error can be zero.

In the case of Pearson uncorrelatedness between two Gaussian random variables sampled via LHS, the minimum correlation error can be zero when N_{sim} = 4l or N_{sim} = 4l+1. Otherwise, the smallest error is of order $1/\Gamma(N_{\text{sim}})$, see a detailed analysis by Vořechovský (2009a).

When a general correlation between two variables is requested, the Spearman correlation is somewhat more difficult to be fulfilled compared to Pearson correlation. The reason is that the number of attainable correlations is much greater in the case of Pearson correlation (it operates with real numbers while Spearman correlation works only with integer ranks).

What deserves attention is the number n_o of so-called *optimal solutions*, i.e. the number of different possible mutual orderings for a given sample size N_{sim} and dimension N_{var} that yields perfect uncorrelatedness. It has been shown (Vořechovský 2009a) that the number of orthogonal Spearman solutions between two random variables is approximately: $n_o \approx 12N_{\text{sim}}^{N_{\text{sim}}-2}/\exp\left(N_{\text{sim}}\right)$, which highlights the rapid increase of n_o with N_{sim}. This result was derived by assuming the density of a random correlation Gaussian – the mode of the Gaussian distribution can be transformed into the number of zero correlation of all the possible $N_{\text{sim}}!$ orderings. Extension of the argumentation into higher dimensions (Vořechovský 2009a) with the assumption that all correlations are independent and identically normally distributed, leads to a number of optimal solutions:

$$n_{o,N_{\text{var}}} \approx \left(N_{\text{sim}}!\right)^{N_{\text{var}}-1}\left(\frac{12}{\sqrt{2\pi}}N_{\text{sim}}{}^{-5/2}\right)^{\binom{N_{\text{var}}}{2}} \tag{6}$$

For a fixed N_{var}, the first factor grows faster with N_{sim} than the second factor decreases and therefore $n_{o,N_{\text{var}}}$ grows with N_{sim}. The main result is the implication that, with increasing sample size N_{sim} the number of optimal solution explodes; the growth is even faster for greater problem dimension N_{var}. Of course, a combination of a very small N_{sim} with large N_{var} may results in nonexistence of an optimal solution.

Analysis of the leading terms in Eq. (6) and postulating that $n_{o,N_{\text{var}}} > 1$ yields a condition on the minimum sample size: $N_{\text{sim}} > 5\ N_{\text{var}}\ /4$.

When orthogonality between two random variables is requested in the sense of zero Pearson correlation, analysis in (Vořechovský 2009a) yields the number

$$n_o \approx 2\sqrt{\pi} N_{\text{sim}}^{\frac{N_{\text{sim}}-1}{2}} \Big/ \exp\left(\frac{N_{\text{sim}}}{2}\right) \tag{7}$$

Comparison of the two results for optimal solution of a pair of variables immediately shows that the number of Spearman orthogonal solutions is approximately equal to a square of Pearson solutions divided by N_{sim}. It is seen that in the Spearman case the number n_o is greater compared number of Pearson orthogonal vectors at the same sample size – orthogonality with real numbers is a stricter requirement than Spearman uncorrelatedness obtained with integer ranks. When the dimension N_{var} gets increased, orthogonal Pearson vectors occur only sparsely.

3.3. Correlation error when $N_{\text{sim}} \leq N_{\text{var}}$

As will be shown later, the performance of the proposed algorithm drastically changes when, at a given problem dimension N_{var}, the sample size exceeds it, i.e. when $N_{\text{var}} > N_{\text{sim}}$. Therefore, it is very important to study the best possible performance for the crossover sample size, i.e. when $N_{\text{sim}} = N_{\text{var}}$.

Every estimated correlation matrix **A** must be positive semidefinite (PSD), i.e. all its eigenvalues are real and nonnegative. It can be shown that when $N = N_{\text{sim}} = N_{\text{var}}$ the correlation matrix **A** is singular (rank deficient) and its rank (number of nonzero eigenvalues) is $N - 1$. The smallest eigenvalue is zero and the determinant of **A**, computed as the product of all eigenvalues, is zero, too. It is also known that the sum of the eigenvalues of **A** equals its trace, which is the sum of the diagonal elements: N_{var}.

A realization of any correlation matrix **A** of order N_{var} can be viewed as a point in N_{c} dimensional space of all N_{c} different correlation coefficients. The volume of the space of all symmetric matrices with off diagonal entries varying from –1 to +1 is V = $2^{N_{\text{c}}}$. It is known that the set of all positive definite matrices is a solid body in that space occupying a region in the vicinity of the origin (a point corresponding to mutual uncorrelatedness). We seek such a realization of **A** (a point in the N_{c}-dimensional space) that is closest to the origin yet represents a singular matrix and therefore remains on the boundary between positive definite matrices and negative definite (invalid) matrices. Mathematical derivation presented by Vořechovský (2009a), based on spectral representation of **A**, yields the lower bound on both ρ_{rms} and ρ_{max}:

$$\underset{N_{\text{var}}=N_{\text{sim}}}{\rho_{\text{rms}}} \geq \frac{1}{N-1}, \quad \underset{N_{\text{var}}=N_{\text{sim}}}{\rho_{\text{max}}} \geq \frac{1}{N-1} \tag{8}$$

These bounds are usually impossible to match in the case of Spearman correlation. The number of attainable correlations is very limited for such a low number of simulations and in dimensions N_{var}>3 there are no correlation matrices with off diagonal elements equal in their absolute values.

The lower bound on correlation errors for cases when $N_{\text{var}} > N_{\text{sim}}$ might have many applications. For example when simulating only a small number of random fields using series expansion methods, see e.g. Vořechovský (2008), the sample size might be much smaller than the number of variables needed for the expansion. There exists other applications when the design of experiments has more dimensions than simulations (supersaturated designs).

Recall that whenever the sample size is smaller than the number of random variables the correlation matrix **A** is singular. The matrix rank can then be computed as: r = rank(**A**) = N_{sim} –1 when ($N_{\mathrm{sim}} \le N_{\mathrm{var}}$) which is, at the same time, the number of non-zero eigenvalues among which the matrix order N_{var} must be distributed. The eigenvalues are uniformly distributed, i.e. the r nonzero eigenvalues are repeated — they equal to N_{var}/($N_{\mathrm{var}} - r$). Analysis if the case of rank=1 yields the best error of one. We conclude by stating that the smallest possible error decreases from one to 1/(N_{sim}–1), and the lower bound reads (full derivation in Vořechovský 2009a):

$$\underset{N_{\mathrm{var}} \ge N_{\mathrm{sim}}}{\rho_{\mathrm{rms}}} \ge \sqrt{\frac{N_{\mathrm{var}} - \left(N_{\mathrm{sim}} - 1\right)}{\left(N_{\mathrm{var}} - 1\right)\left(N_{\mathrm{sim}} - 1\right)}} \tag{9}$$

The lower bound on ρ_{max} is more complicated (Vořechovský 2009a).

4. Results of performance tests

4.1. General remarks

From here on, we present results of large numerical studies of performance of the technique proposed by (Vořechovský 2002, 2007) and (Vořechovský and Novák 2002, 2003, 2009). We compare the results with both the theoretical bounds presented above and performance of the two other known algorithms. In particular the ability to fulfill the desired correlation structure while keeping the sampled marginal distributions intact is measured for both (i) *uncorrelated* and (ii) strongly *correlated* cases. We have conducted estimates of statistics of errors (1) ρ_{rms} and (2) ρ_{max} for a vast number of combinations of N_{sim} and N_{var}. These tests were carried out for both (a) Spearman and (b) Pearson correlation coefficient (in the latter case the margins were Gaussian and LHS-sampled).

Both errors ρ_{rms} and ρ_{max} can be viewed as random variables because they are results of stochastic optimization that depends on random starting conditions – the process of correlation control depends on the seed of computer (pseudo) random number generator. Therefore, in order to deliver statistically significant estimations on the random performance we conducted a large number of runs with the same input settings. The number of runs is not a constant – for large sample sizes the variability of results is small compared to a small sample size, where we used 200 runs. In each run we used the same starting conditions except for the seed of random number generator. The recorded statistics are the sample means and standard deviations and the recorded minima and maxima. In this paper, the performance is always presented as the correlation error ρ_{rms} or ρ_{max} versus the sample size N_{sim} in a logarithmic plot. The reason is that the graphs contain plots of power laws that appear as straight lines.

The algorithm is compared to two previously developed and well known techniques for correlation control: Owen's (1994) method based on Gram-Schmidt orthogonalization procedure for generation of uncorrelated vectors and (b) Iman and Conover's (1980) method based on Cholesky decomposition of correlation matrix. Comparison is performed only for ρ_{rms} which is somewhat more practical error measure. Average performance of Owen's method is denoted as $\rho_{\mathrm{rms}}^{\mathrm{RGS}}$ and average performance of Iman and Conover's method is $\rho_{\mathrm{rms}}^{\mathrm{RC}}$. The performance of our algorithm is denoted as $\rho_{\mathrm{rms}}^{\mathrm{SA}}$.

4.2. TESTS WITH UNCORRELATED VARIABLES

We start with performance graphs for target uncorrelatedness, i.e the target correlation **T**=**I** (unit matrix). This is the most frequently used case in Monte Carlo simulation approaches.

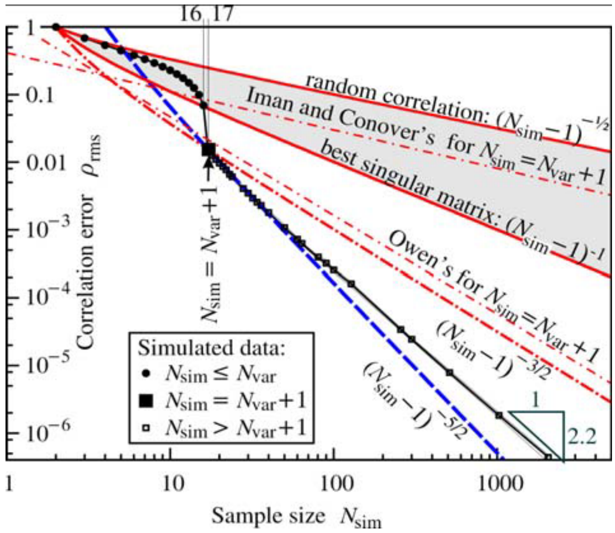


Figure 2. Typical performance plot obtained for $N_{\text{var}} = 16$. Simulated data (averages are denoted by symbols and a solid line, minima and maxima by thin solid lines) are compared to theoretical bounds and results of other techniques.

Before we present a plot of correlation performance for all tested N_{var} in a single plot, we show and comment a typical performance plot together with all theoretically derived bounds and comparisons to previously developed algorithms in Fig. 2. We have selected the dimension of the problem to be $N_{\text{var}} = 16$ variables and performed simulations for $N_{\text{sim}}=2,\ldots,1000$.

First of all, the average performance has *two distinct branches*: (i) graph for $N_{\text{sim}} \leq N_{\text{var}}$ that appear nonlinear in the plot and (ii) second branch which is approximately a *power law* when $N_{\text{sim}} > N_{\text{var}}$. In the former case, the correlation matrices are singular (at least one zero eigenvalue). There is a clear boundary between the two regimes that corresponds to $N_{\text{sim}} = N_{\text{var}}$ (see the line denoted as "best singular matrix" in Fig. 2), compare with Eq. (8):

$$\rho_{\text{rms}}^{\text{SA}} = 1/\left(N_{\text{sim}} - 1\right) \approx N_{\text{sim}}^{-1} \tag{10}$$

Results of our SA algorithm for the *first branch* ($N_{\text{sim}} \leq N_{\text{var}}$) has no variability, i.e. the solution was identically good for any seed setting (run) and it almost perfectly follows Eq. (9). The error $\rho_{\text{rms}}^{\text{SA}}$ never exceeds the upper bound given by a random ordering correlation (Eqs. 4 and 5), i.e. $\approx N_{\text{sim}}^{-1/2}$ (see the line denoted as "random correlation"). We can say that whenever $N_{\text{sim}} \leq N_{\text{var}}$, the $\rho_{\text{rms}}^{\text{SA}}$ error stays between the two power laws, i.e. within the shaded area in Fig. 2. This holds for an arbitrary N_{var}, not only sixteen. The situation for the other norm $\rho_{\max}^{\text{SA}}$, is almost identical. The only difference is that the upper bound is somewhat greater (but known and never violated, see (Vořechovský 2009a and 2009b) for details) and the lower bound is somewhat higher, too.

Results of our SA algorithm for the *second branch* ($N_{\mathrm{sim}} > N_{\mathrm{var}}$) are very interesting. First of all, when $N_{\mathrm{sim}} = N_{\mathrm{var}} + 1$, the error $\rho_{\mathrm{rms}}^{\mathrm{SA}}$ drastically decreases (from the one for $N_{\mathrm{sim}} = N_{\mathrm{var}}$) because the correlation matrix becomes positive definite, see the sudden jump when N_{sim} changes from 16 to 17 in Fig. 2 (jump from the solid circle to the solid box). However, the error starts to have a certain scatter showing that the algorithm stops with differently good solutions depending on the seed settings. It will be shown that the average error for $N_{\mathrm{sim}} = N_{\mathrm{var}} + 1$ is almost exactly a power law as follows (see the thick dash-dot line):

$$\underset{N_{\mathrm{sim}}=N_{\mathrm{var}}+1}{\rho_{\mathrm{rms}}^{\mathrm{SA}}} = \frac{1}{\left(N_{\mathrm{sim}}-1\right)^{3/2}} \approx N_{\mathrm{sim}}^{-1.5} \tag{11}$$

As the sample size N_{sim} increases (for a given N_{var}), the average error appears as a *straight line* with an universal slope of almost –5/2 and therefore the average error reaches almost exactly (dashed line):

$$\underset{N_{\mathrm{sim}}>N_{\mathrm{var}}}{\rho_{\mathrm{rms}}^{\mathrm{SA}}} = N_{\mathrm{var}} \frac{1}{\left(N_{\mathrm{sim}}-1\right)^{5/2}} \approx N_{\mathrm{var}} N_{\mathrm{sim}}^{-2.5} \tag{12}$$

One can easily check that by substituting $N_{\mathrm{sim}} = N_{\mathrm{var}} + 1$ into Eq. (12) the formula in Eq. (11) gets recovered. In fact the slope of –5/2 (discussed by Vořechovský 2009a and 2009b) is not exactly reached for the whole tested spectrum of N_{sim}. Numerical simulations with varied N_{trials} as (a parameter in the SA algorithm) suggest that the optimal power of average convergence (–5/2) is reached only for a sufficiently large number of iterations N_{trials}. As the sample size increases, the demands on N_{trials} grow. For a constant N_{trials} that does not reflect the "size" of the problem (number of possible rank combinations) the power changes from –2.5 to approximately –2.2, which is a somewhat worse performance (see the triangle at bottom right in Figs. 2 and 3).

Let us compare the performance of our SA algorithm with Owen's $\rho_{\mathrm{rms}}^{\mathrm{RGS}}$ and Iman and Conover's $\rho_{\mathrm{rms}}^{\mathrm{RC}}$. For simplicity, the comparison is made only for average errors and only for the particular situation of $N_{\mathrm{sim}} = N_{\mathrm{var}} + 1$. In fact, we are ready to compare the three dash-dot lines in Fig. 2.

As reported by Owen (1994), the average errors obtained by ordinary regressions of data obtained by the known algorithms give:

$$\begin{aligned} \rho_{\mathrm{rms}}^{\mathrm{RC}} &\approx +0.42 N_{\mathrm{sim}}^{-0.57} \quad (\text{Iman and Conover}) \\ \rho_{\mathrm{rms}}^{\mathrm{RGS}} &\approx +1.35 N_{\mathrm{sim}}^{-1.45} \quad (\text{Owen}) \end{aligned} \tag{13}$$

One can immediately see that Iman and Conover's algorithm has much worse performance because the slope (–0.57) is significantly milder than Owen's –1.45. Owen's algorithm has almost the same slope as ours –1.5 (see Eq. 11) but the plot is somewhat shifted towards 35% grater errors (see the constant of 1.35 and compare the lines in Fig. 2).

Unfortunately, when Owen's algorithm is used for $N_{\mathrm{sim}} > N_{\mathrm{var}} + 1$, the convergence rate decreases (!) to $\rho_{\mathrm{rms}}^{\mathrm{RGS}} \propto N_{\mathrm{sim}}^{-1}$ which might seem to be strange. The implication of this is that adding more simulations to the problem makes the situation worse. In other words, it is better to generate correlations for a greater N_{var} than needed and remove the unnecessary variables afterwards. Owen (1994) gives an explanation to this fact.

Note that in our SA case, *increase of the sample size* for a given N_{var} results in a *drastic improvement of the correlation error* (Eq. 12) which is logical: a greater sample size allows for many more combinations of ranks to select a sub-optimal ordering from.

Regarding a random scatter of our performance results: as mentioned before, we only have some scatter when $N_{\mathrm{sim}} > N_{\mathrm{var}}$ (second branch). It can be seen from Fig. 3 that for a constant N_{var}, the scatter band is equally wide for various sample sizes. This suggests that the coefficient of variation of $\rho_{\mathrm{rms}}^{\mathrm{SA}}$ is a constant, independent of N_{sim}. Indeed, the standard deviation of $\rho_{\mathrm{rms}}^{\mathrm{SA}}$ is almost exactly equal to $N_{\mathrm{sim}}^{-5/2}$ (negligible differences are found in regressions of the data). Therefore, the coefficient of variation is inversely proportional to the number of variables:

$$\underset{N_{\mathrm{sim}}>N_{\mathrm{var}}}{\mathrm{cov}\left[\rho_{\mathrm{rms}}^{\mathrm{SA}}\right]} \approx \frac{N_{\mathrm{sim}}^{-5/2}}{N_{\mathrm{var}} N_{\mathrm{sim}}^{-5/2}} = \frac{1}{N_{\mathrm{var}}} \tag{14}$$

This corresponds to the decay of scatter band width for growing N_{var} as seen in Fig. 3.

Note that performance graphs for Spearman and Pearson correlations are almost identical (differences are found only for very small N_{var}, a fact thoroughly explained in Vořechovský 2009a).

Also, we can say that the overall shape of convergence graphs looks similar for $\rho_{\mathrm{max}}^{\mathrm{SA}}$ and $\rho_{\mathrm{rms}}^{\mathrm{SA}}$ (the former norm is more conservative and the curves are shifted upwards). Detailed analyses of the differences between the two norms can be found in (Vořechovský 2009a and 2009b).

4.3. TESTS WITH CORRELATED VARIABLES

In order to study the algorithm performance for *correlated variables*, we have to select some correlation matrix pattern because there is no unique one (as opposed to the uncorrelatedness). The target correlation matrix was constructed (i) to cover a spectrum of target correlation coefficients from interval $T_{i,j} \in \langle -1,1 \rangle$ and (ii) to be surely positive definite. In particular, each entry of the target correlation matrix **T** was a product of two numbers associated with *i*th and *j*th variable:

$$T_{i,j} = T_i \, T_j, \quad i, j = 1, \ldots, N_{\mathrm{var}} \tag{15}$$

where $T_i = 2(i\text{-}1)/ N_{\mathrm{var}} - 1$ (uniform distribution over interval (–1;1)). Such a separable product correlation structure ensures the symmetry and positive definiteness of **T**. The patterns of correlation matrices for arbitrary N_{var} look similar. The minimum off-diagonal correlation coefficient in **T** reads $T_{1,N_{\mathrm{var}}} = 2/N_{\mathrm{var}} - 1$ and the maximum one is $T_{1,2} = 1 - 2/N_{\mathrm{var}}$. There are always $N_{\mathrm{var}} - 1$ pairs of variables with zero target correlation coefficients in **T**. In Fig. 3bottom left, there is an image of the absolute values of correlation matrix **T** in the bottom left corner and the figure next to it shows the distribution of correlations $T_{i,j}$ ranging from –1 to 1.

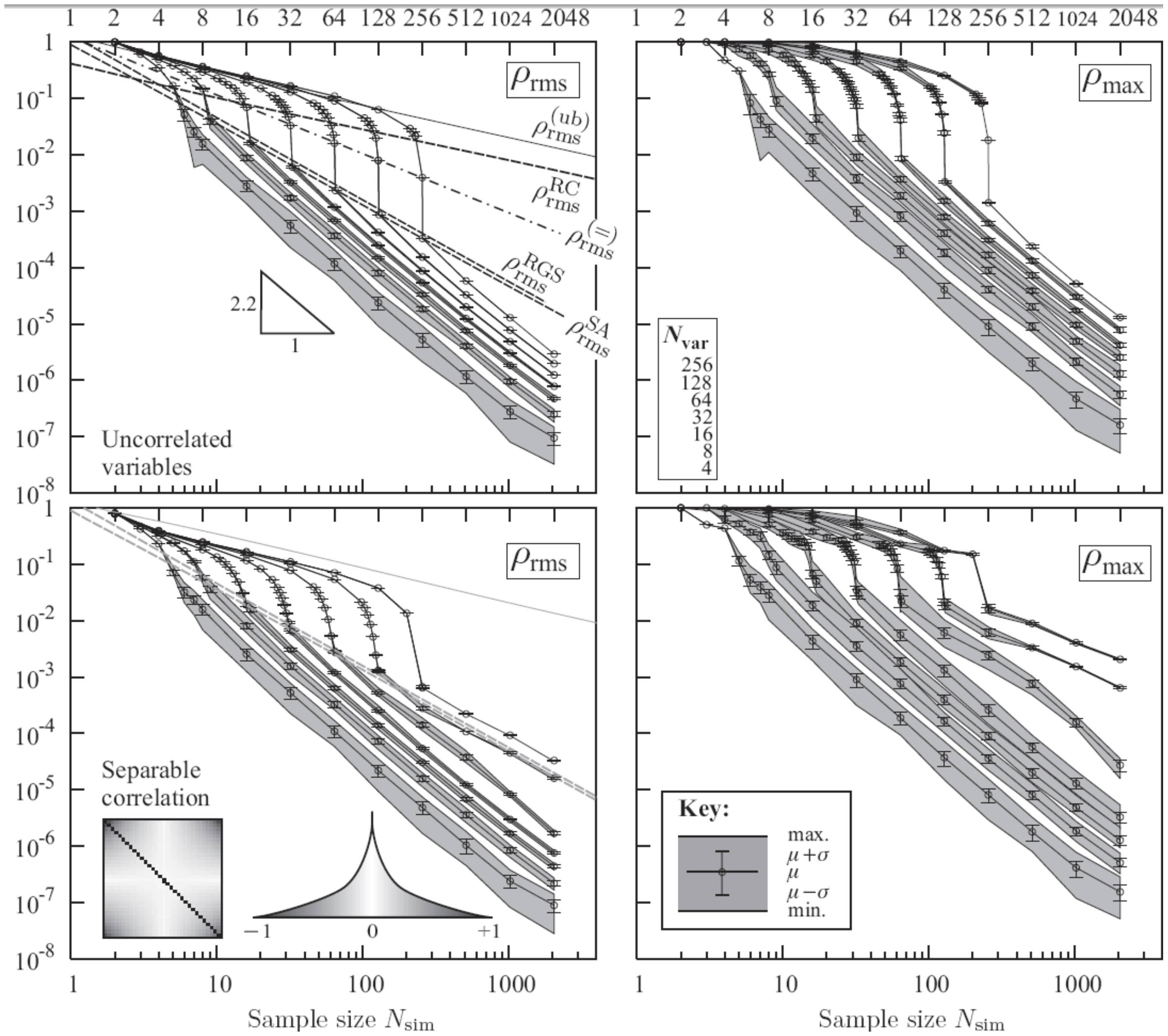


Figure 3. Results of performance study. Top: Uncorrelated variables; Bottom: Separable correlation. Left: ρ_{rms}; Right: ρ_{max}. Bottom left: Image of correlation matrix (absolute values of $T_{\mathrm{i,j}}$ and distribution of correlation coefficients in **T** for the case of studied separable correlation).

Generally, the correlated test was a tough task for the algorithm, as compared to the uncorrelated case. The results plotted in the bottom line can hardly be distinguished from the result in top of the figure for small problems. For large sample sizes, however, the number of trials must be large enough to deliver correlation errors as good as in the uncorrelated case. The reason is that uncorrelatedness is the most frequent pattern, while “higher correlatedness” can be matched by a smaller number of mutual rankings of samples.

4.4. DISCUSSION AND REMARKS

In most cases the solution found by the algorithm is far too good in terms of the correlation error. The usual requirements on the correlation are not that strict. In practice, the information on the dependency is expressed just as a correlation, very often only vaguely described as “strong” or “weak”. The computational time gets very short when stopping condition in a form of the error tolerance is set.

Clearly, the algorithm is designed to seek configurations with small values of ρ_{rms} (and consequently ρ_{max}). However, it might be questioned what solution it yields in terms of the dependence pattern of the desired random vector. We know that the information provided to the algorithm (marginals and correlation structure) does not suffice for constructing a unique joint probability density function (jpdf). This fact represents a possible danger: the generated samples may represent some unwanted jpdf. This is discussed by Vořechovský (2009b). Regarding the solutions with small errors ρ_{max} and ρ_{rms}: it must be questioned whether such a requirement is a good one. Especially in the case of uncorrelated variables there exists a risk of obtaining a sample with unwanted and strange patterns of dependence. A simple remedy is suggested via enhancing the cost function with either canonical correlations or copulas (or at least: combination of different kinds of correlation measures such as Pearson and Spearman). Such an extension is simple Vořechovský (2009b). In this way, the algorithm shows the non-uniquness of some of the reliability approaches based on Rosenblat (1952) transformation (see e.g. Hohenbichler and Rackwitz 1981) or Nataf (1962) transformation (promoted by Liu and Kiureghian 1986) known also as Li-Hammond (1975) or NORTA transformation.

5. Conclusions

The paper promotes a number of innovative results on correlation bounds based on combinatorial analysis of the problem, linear algebra and theory of probability. These bounds are compared to results of suggested algorithm for correlation control based on stochastic combinatorial optimization. The results support a statement that the proposed algorithm operates with results very close to the lower bounds on performance. The algorithm is considerably more efficient than the other two well-known algorithms, i.e. Iman and Conover’s (1980) Cholesky decomposition and Owen’s (1994) Gram-Schmidt orthogonalization.

Acknowledgements

The presented work was supported by The Grant Agency of the Academy of Sciences of the Czech Republic under project KJB201720902 and partly also by the Ministry of Education, Youth and Sports of the Czech Republic under project MSMT 0021630519.

References

Conover, W.J. On a Better Method for Selecting Input Variables, unpublished Los Alamos National Laboratories manuscript, reproduced as Appendix A of "*Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems*" by J.C. Helton and F.J. Davis, *Sandia National Laboratories report* SAND2001-0417, printed November 2002. 1975

Ghosh, S., Henderson, S. G. Behavior of the NORTA method for correlated random vector generation as the dimension increases, *ACM Transactions on Modeling and Computer Simulation* 13 (3), 276–294, 2003.

Hohenbichler, M. Rackwitz, R. Non-normal dependent vectors in structural safety, *Journal of Engineering Mechanics*, ASCE 107 (6), 1227–1238, 1981.

Hotelling, H. and M.R. Pabst, Rank correlation and tests of significance involving no assumption of normality, *The Annals of Mathematical Statistics* 7 (1), 29–43, 1936.

Iman, R. C. and W. J. Conover, Small sample sensitivity analysis techniques for computer models with an application to risk assessment, *Communications in Statistics: Theory and Methods* A9 (17), 1749–1842, 1980.

Iman, R. C., and W. J. Conover. A distribution free approach to inducing rank correlation among input variables, *Communications in Statistics* B11, 311–334, 1982.

Johnson, M. E. *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*, Wiley Series In Probability And Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1987.

Li, S.T. and J.L. Hammond. Generation of pseudo-random numbers with specified univariate distributions and correlation coefficients, *IEEE Transactions on Systems, Man, Cybernetics* 5, pp. 557–560, 1975.

Liu, P. and A. Der Kiureghian. Multivariate distribution models with prescribed marginals and covariances, *Probabilistic Engineering Mechanics* 1 (2) (1986) 105–111, 1986.

McKay, M. D. and W.J. Conover. and R.J. Beckman, A comparison of three methods for select-ing values of input variables in the analysis of output from a computer code, *Technometrics*, 21, 239–245, 1979.

Metropolis, N. and A.W. Rosenbluth and M.N. Rosenbluth and A.H. Teller and E. Teller. Equation of state calculations by fast computing machines, *Journal of Chemical Physics* 21, pp. 1087– 1092, ISSN: 0021-9606, 1953.

Nataf, A. Détermination des distributions de probabilités dont les marges sont donnés, *Comptes Rendus de L'Académie des Sciences* 225, 42–43, 1962.

Owen, A. B. Controlling Correlations in Latin Hypercube Samples, *Journal of the American Statistical Association (Theory and methods)* 89 (428), 1517–1522, 1994.

Pearson, K. On further methods of determining correlation, *Journal of the Royal Statistical Society* 70 (4) 655–656, 1907.

Rosenblatt, M. Remarks on multivariate analysis, *Annals of Statistics* 23, 470–472, 1952.

Vořechovský, M. Performance study of correlation control in Monte Carlo type simulation. In 3rd *PhD Workshop Brno-Prague-Weimar*, also ISM-Bericht 1/2006 Bauhaus-Univeristät Weimar, Weimar, Germany, pp. 35-38, 2006.

Vořechovský, M. *Stochastic computational mechanics of quasibrittle structures*. Habilitation thesis presented at Brno University of Technology, Brno, Czech Republic, 2007.

Vořechovský, M. Nové úpravy simulační metody Latin Hypercube Sampling a možnosti využití (New improvements to simulation technique Latin Hypercube Sampling and possibilities of its utilization). In M. In. Stibor (Ed) Problémy modelování (Problems of Modeling) conference, Faculty of Civil Engineering VŠB-TUO, Ostrava, Czech Republic, 2002. Brno University of Technology, pp. 83-90, in Czech, 2002.

Vořechovský, M. Simulation of simply cross correlated random fields by series expansion methods, *Structural safety* (Elsevier) 30 (4), 337—363, 2008.

Vořechovský, M., Correlation control in small sample Monte Carlo type simulations II: Theoretical analysis and performance bounds, *Probabilistic Engineering Mechanics* (Elsevier), in review, 2009a.

Vořechovský, M., Correlation control in small sample Monte Carlo type simulations III: Algorithm performance and relevance to copulas and multivariate modeling, *Probabilistic Engineering Mechanics* (Elsevier), in review, 2009b.

Vořechovský, M., Novák, D. Correlated random variables in probabilistic simulation. In: Schießl, P. et al. (Eds.), 4th *International Ph.D. Symposium in Civil Engineering*. Vol. 2. Millpress, Rotterdam. Munich, Germany, pp. 410–417. Awarded paper. ISBN 3-935065-09-4, 2002.

Vořechovský, M. and D. Novák. Statistical correlation in stratified sampling. In: Der Kiureghian et al. (Eds.), ICASP 9, *International Conference on Applications of Statistics and Probability in Civil Engineering*. Millpress, Rotterdam. San Francisco, USA, pp. 119–124. ISBN 90 5966 004 8, 2003.

Vořechovský, M. and D. Novák. Correlation control in small sample Monte Carlo type simulations I: A simulated annealing approach, *Probabilistic Engineering Mechanics (Elsevier)*, 24(3):452-462, 2009.