Probabilistic Engineering Mechanics 29 (2012) 105–120

Contents lists available at SciVerse ScienceDirect

Probabilistic Engineering Mechanics

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

Correlation control in small sample Monte Carlo type simulations II: Analysis of estimation formulas, random correlation and perfect uncorrelatedness

M. Vořechovský *

Institute of Structural Mechanics, Brno University of Technology, Veveří 95, 602 00 Brno, Czech Republic

ARTICLE INFO

Article history:
Received 28 April 2011
Accepted 26 September 2011
Available online 2 October 2011

Keywords:
Statistical correlation distribution
Pearson correlation
Spearman correlation
Kendall correlation
Uncorrelatedness
Random correlation
Orthogonal vectors
Multivariate Monte Carlo simulation
Latin hypercube sampling

ABSTRACT

This paper presents a number of theoretical and numerical results regarding correlation coefficients and two norms of correlation matrices in relation to correlation control in Monte Carlo type sampling and the designs of experiments. The paper studies estimation formulas for Pearson linear, Spearman and Kendall rank-order correlation coefficients and formulates the lower bounds on the performance of correlation control techniques such as the one presented in the companion paper Part I. In particular, probabilistic distributions of the two norms of correlation matrices defined in Part I are delivered for an arbitrary sample size and number of random variables in the case when the sampled values are ordered randomly. Next, an approximate number of designs with perfect uncorrelatedness is estimated based on the distribution of random correlation coefficients. It is shown that a large number of designs exist that perfectly match the unit correlation matrix.

1. Introduction

Statistical sampling is of interest not only to statisticians, but to a variety of research fields such as engineering, economics, design of experiments, and operational research. Analysts are often faced with time-consuming and expensive sampling (physical or virtual) to achieve statistically significant output statistics. When sampling from random vectors, it is important to control correlations or even dependence patterns between marginals.

In Part I [1] the correlation control problem was approached from an *optimization* perspective. We have developed a combinatorial optimization algorithm based on Simulated Annealing to shuffle the realizations in the sampling plan in order to achieve a good match between the target and estimated correlation matrices.

When applying any correlation control technique, it is important to know what the bounds on the correlation errors are. This information can be useful e.g. for the selection of stopping criteria in algorithms employed for correlation control. In order to quantify an error in the correlations of a given sample, one must (i) select a correlation estimator and (ii) define a scalar measure of the correlation matrix. Regarding the first item, two point estimators are among those most widely utilized, namely the linear Pearson sample correlation and the rank-order Spearman and Kendall estimators. When dealing with a multivariate case, a matrix consisting of differences between the target correlation coefficients and the estimated ones was defined in Part I [1] together with two matrix norms: $\rho_{\max}$ and ρ_{rms}. These norms represent a natural choice and they are used in the assessment of correlation errors; see e.g. [2–4].

When a suitable error matrix norm is defined, an analyst preparing a sample is interested in achieving the lowest possible value of that norm. This paper analyzes various situations for which the bounds on the correlation errors $\rho_{\max}$ and ρ_{rms} are derived exactly or at least estimated.

Another goal of controlled statistical sampling is usually to perform the sampling with the smallest possible sample size, N_{sim}, and yet achieve statistically significant estimates of the response. One of the most well-known variance-reduction sampling techniques of the Monte Carlo type is Latin hypercube sampling (LHS). This sampling technique will be considered in some parts of this paper in conjunction with the correlation between Gaussian-distributed random variables.

The paper is organized as follows. Section 2 analyzes the formulas for sample correlation coefficients and Section 3 provides some results in terms of the number of attainable values of estimated correlation. Section 4 analyzes the correlation emerging from samples with randomly permuted values. Based on the probabilistic distribution of correlation, the distributions of errors $\rho_{\max}$ and ρ_{rms} are derived for the arbitrary sample size N_{sim} and vector dimension N_{var}. Section 5 builds on these results and estimates the numbers of sample permutations that yield perfect

* Tel.: +420 54114 7370; fax: +420 541240994.
E-mail address: vorechovsky.m@fce.vutbr.cz.
URL: http://www.fce.vutbr.cz/STM/vorechovsky.m/.

0266-8920/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.probengmech.2011.09.004

uncorrelatedness (which is very often requested when analyzing problems with independent inputs).

2. Revision of correlation estimation

The estimated correlation matrix is a symmetric matrix of the order N_{var} and can be written as the sum

$$\boldsymbol{A} = \mathbf{I} + \boldsymbol{L} + \boldsymbol{L}^T \tag{1}$$

where $\mathbf{I}$ is the identity matrix and $\boldsymbol{L}$ is the strictly lower triangular matrix with entries within the range $\langle -1, 1\rangle$. There are N_c correlations (entries in the $\boldsymbol{L}$ matrix) that describe pairwise correlations:

$$N_c = \binom{N_{\text{var}}}{2} = \frac{N_{\text{var}}\left(N_{\text{var}} - 1\right)}{2}. \tag{2}$$

2.1. Norms of correlation error

Eqs. (11) and (14) from the companion Part I [1] define the norms ρ_{rms} and $\rho_{\max}$ of the error matrix $\boldsymbol{E}$, a matrix obtained as the difference between the target ($\boldsymbol{T}$) and the actual (estimated, $\boldsymbol{A}$) correlation matrices. In this section, the target correlation matrix is the unit matrix $\boldsymbol{T} = \mathbf{I}$ (uncorrelated variables are targeted) and therefore the error norms can be reformulated as norms of the actual correlation matrix $\boldsymbol{A}$. Moreover, we ignore the weighting of various entries in the correlation matrix introduced in [1,5] (or, we in fact consider unit weights) and the two norms can be written as follows. The *absolute* norm reads:

$$\rho_{\max} = \max_{1\le i<j\le N_{\text{var}}} \left|A_{i,j}\right| \tag{3}$$

and the *root mean square* correlation error

$$\rho_{\text{rms}} = \sqrt{\frac{1}{N_c}\sum_{i=1}^{N_{\text{var}}-1}\sum_{j=i+1}^{N_{\text{var}}} A_{i,j}^2}, \tag{4}$$

i.e. the square root of the average of the squares of all off-diagonal correlation matrix entries.

We now review the sampling versions (estimators) of the three most frequently used correlation coefficients.

2.2. Pearson correlation coefficient (sampling)

The most well-known correlation measure is the linear Pearson correlation coefficient (PCC). The PCC takes values from between -1 and $+1$, inclusive, and provides a measure of the strength of the linear relationship between two variables. The actual PCC between two variables, say X_i and X_j, is estimated using the sample correlation coefficient A_{ij} as

$$A_{ij} = \frac{\sum_{s=1}^{N_{\text{sim}}}\left(x_{i,s} - \overline{X_i}\right)\left(x_{j,s} - \overline{X_j}\right)}{\sqrt{\sum_{s=1}^{N_{\text{sim}}}\left(x_{i,s} - \overline{X_i}\right)^2 \sum_{s=1}^{N_{\text{sim}}}\left(x_{j,s} - \overline{X_j}\right)^2}}, \tag{5}$$

$$\overline{X_i} = \frac{1}{N_{\text{sim}}}\sum_{s=1}^{N_{\text{sim}}} x_{i,s} \qquad \overline{X_j} = \frac{1}{N_{\text{sim}}}\sum_{s=1}^{N_{\text{sim}}} x_{j,s}.$$

When the actual data $x_{i,s}$, $s = 1, 2, \ldots, N_{\text{sim}}$ of each vector $i = 1, 2, \ldots, N_{\text{var}}$ are standardized into $z_{i,s}$, i.e. into vectors that yield zero average and unit sample variance estimates, the formula simplifies to

$$A_{ij} = \frac{1}{N_{\text{sim}}}\sum_{s=1}^{N_{\text{sim}}} z_{i,s}\, z_{j,s} \tag{6}$$

which is the dot product (or scalar product) of two vectors divided by the sample size.

2.3. Spearman correlation coefficient (sampling)

The formula for Spearman (nonparametric or distribution-free) correlation coefficient estimation is identical to the one for Pearson linear correlation with the exception that the values of random variables X_i and X_j are replaced with the ranks $\pi_{i,s}$ and $\pi_{j,s}$, $s = 1, \ldots, N_{\text{sim}}$. The ranks are permutations of numbers, s. It is convenient to transform the ranks into $r_{i,s} = \pi_{s,i} - \overline{\pi}_i$ and $r_{j,s} = \pi_{s,j} - \overline{\pi}_j$ where

$$\overline{\pi}_i = \overline{\pi}_j = \overline{\pi} = \frac{1}{N_{\text{sim}}}\sum_{s=1}^{N_{\text{sim}}} s = \frac{N_{\text{sim}} + 1}{2} \tag{7}$$

is the average rank. The (actual) rank correlation is then defined as

$$A_{ij} = \frac{\sum r_{i,s}\, r_{j,s}}{\sqrt{\sum r_{i,s}^2 \sum r_{j,s}^2}}, \tag{8}$$

the sums being over the N_{sim} values in the sample. By noting that the sum of the first N_{sim} squared integers is $N_{\text{sim}}\left(N_{\text{sim}} + 1\right)\left(2N_{\text{sim}} + 1\right)/6$, we find that $\sum r_{j,s}^2 = \sum r_{j,s}^2 = (N_{\text{sim}}^3 - N_{\text{sim}})/12$, and the rank correlation reads:

$$A_{ij} = \frac{12\sum r_{i,s}\, r_{j,s}}{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)} = \frac{12\sum \pi_{i,s}\,\pi_{j,s}}{N_{\text{sim}}^3 - N_{\text{sim}}} - 3\frac{N_{\text{sim}} + 1}{N_{\text{sim}} - 1}. \tag{9}$$

In the case of ties (identical ranks for pairs of values of a single variable), the averaged ranks are used. Note that when LHS is applied to continuous parametric distributions no ties can occur in the generated data. Therefore, we do not consider ties from here on. Another formula exists for Spearman correlation suitable for data with no ties. The (actual) correlation coefficient between any two vectors each consisting of permutations of integer ranks from 1 through N_{sim} is

$$A_{i,j} = 1 - \frac{6D}{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)} \tag{10}$$

where D is the sum of values d_s, i.e. the differences between the sth integer elements in the vectors:

$$D = \sum_{s=1}^{N_{\text{sim}}} d_s^2. \tag{11}$$

Every mutual permutation of ranks can be achieved by permuting the ranks π_s of the second variable against the identity permutation corresponding to the ranks of the first variable. Therefore, we may write

$$D = \sum_{s=1}^{N_{\text{sim}}}(s - \pi_s)^2 = 2\left[\sum_{s=1}^{N_{\text{sim}}} s^2 - \sum_{s=1}^{N_{\text{sim}}}(s\pi_s)\right]. \tag{12}$$

This is equal to $N_{\text{sim}}\left(N_{\text{sim}} + 1\right)\left(2N_{\text{sim}} + 1\right)/3 - 2\sum(s\pi_s)$, revealing the equivalence between Eqs. (9) and (10).

Spearman correlation can, in general, take any value between -1 and $+1$, inclusive, depending on the value of the sum $\sum d_s^2$. The lowest correlation (perfect negative dependence) is achieved for the reverse ordering of rank numbers ($\pi_s = N - s + 1$) and corresponds to the case when the sum D equals $N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)/3$. Conversely, the maximum correlation (perfect positive dependence) is achieved for identical ranks ($\pi_s = s$) and the sum equals zero. That is why

$$D \in \langle 0; u_N\rangle, \quad \text{where } u_N = \frac{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)}{3}. \tag{13}$$

2.4. Kendall correlation coefficient (sampling)

Kendall's [6] (nonparametric or distribution-free) correlation coefficient estimates the difference between the probability of concordance and discordance between two variables, x_i and x_j. For data without ties, the estimate is calculated based on the rankings π_i and π_j of N_{sim} samples of two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Let us index the ranks (or samples) by $1 \leq k, l \leq N_{\text{sim}}$. The formula for sample correlation is a direct estimation of the difference between the probabilities:

$$A_{ij} = \frac{n_c - n_d}{\binom{N_{\text{sim}}}{2}} = \frac{\sum\limits_{k<l}^{N_{\text{sim}}} \operatorname{sgn}\left[\left(\pi_{i,k} - \pi_{i,l}\right)\left(\pi_{j,k} - \pi_{j,l}\right)\right]}{\binom{N_{\text{sim}}}{2}} \tag{14}$$

where $\operatorname{sgn}(z) = -1$ for negative z, $+1$ for positive z, and zero for $z = 0$. The numerator counts the difference between concordant pairs n_c and discordant pairs n_d. The denominator is the maximum number of pairs with the same order, i.e. the total number of item pairs with respect to which the rankings can be compared. The number of concordant pairs n_c is the number of item pairs on the order of which both rankings agree, i.e. a pair $\left(\pi_{i,k}, \pi_{j,k}\right)$ and $\left(\pi_{i,l}, \pi_{j,l}\right)$ of points in the sample is concordant if either $\pi_{i,k} < \pi_{i,l}$ and $\pi_{j,k} < \pi_{j,l}$ or, $\pi_{i,k} > \pi_{i,l}$ and $\pi_{j,k} > \pi_{j,l}$. Analogically, n_d is the number on which both rankings disagree.

The number of concordant pairs can be calculated by adding scores: (i) a score of one for every pair of objects that are ranked in the same order and (ii) a zero score for every pair that are ranked in different orders:

$$n_c = \sum_{k=1}^{N_{\text{sim}}-1} \sum_{l=k+1}^{N_{\text{sim}}} \left(\mathbf{1}_{\left(\pi_{i,k}-\pi_{i,l}\right)\left(\pi_{j,k}-\pi_{j,l}\right)>0}\right) \tag{15}$$

where the indicator function $\mathbf{1}_A$ equals one if A is true, and zero otherwise. Analogically, n_d would count only for opposite orders and the formula would be identical but with opposite orientation of the inequality sign.

In the case of tied rank, the denominator is usually adjusted. We do not consider ties from here on. Therefore, the above Eq. (14) can be rewritten by exploiting the fact that the number of pairs is the sum of concordant and discordant pairs and therefore the number of discordant pairs is $n_d = \binom{N_{\text{sim}}}{2} - n_c$. This can be substituted into Eq. (14), yielding

$$A_{ij} = \frac{4\, n_c}{N_{\text{sim}}\left(N_{\text{sim}} - 1\right)} - 1 = 1 - \frac{4\, n_d}{N_{\text{sim}}\left(N_{\text{sim}} - 1\right)}. \tag{16}$$

A straightforward implementation of the algorithm based on the above equations has $\mathcal{O}\left(N_{\text{sim}}^2\right)$ complexity. In practise, it is convenient to rearrange the two rank vectors so that the first one is in increasing order. As proposed in [7], a more sophisticated algorithm based on the Merge Sort algorithm can then be employed to compute the coefficient in $\mathcal{O}\left(N_{\text{sim}} \cdot \log N_{\text{sim}}\right)$ time.

Kendall's correlation coefficient is intuitively simple to interpret. When compared to the Spearman coefficient, its algebraic structure is much simpler. Note that Spearman's coefficient involves concordance relationships among three sets of observations, which makes the interpretation somewhat more complex than that for Kendall's coefficient. Regarding the relation between Spearman's correlation (say ρ) and Kendall's correlation (say τ), the bounds of Daniel's [8] universal inequality $|3\tau - 2\rho| \leq 1$ have been independently refined in [9,10]:

$$\tau - \left(1 - \tau^2\right) \leq 3\tau - 2\rho \leq \tau + \left(1 - \tau^2\right). \tag{17}$$

A simple proof of these bounds has recently been given in [11].

Kendall's correlation can, in general, take any value between -1 and $+1$, inclusive.

For many joint distributions, Spearman's rho and Kendall's tau have different values, as they measure different aspects of the dependence structure [12]. It has long been known about the relationship between the two measures that, for many distributions exhibiting weak dependence, the sample value of Spearman's rho is about 50% larger than the sample value of Kendall's tau [13,14]. Indeed, the ratio between the population versions of rho and tau has recently been shown to approach (as $N_{\text{sim}} \to \infty$) the limiting ratio of 3/2 as the joint distribution approaches that of two independent variables and [15]. The same ratio has also been proved to hold for extreme order statistics [16,17].

3. Theoretical bounds on estimated correlation errors

This section analyzes the theoretical bounds of correlation control algorithm performance for a pair of random variables. We focus on three correlation coefficients: Pearson, Spearman and Kendall's correlations. The most detailed results in the remainder of this section and also in this paper will be stated for the Spearman and Kendall rank correlation. This is because these correlation measures are nonparametric measures and they are distribution-free; the results for these widely-used coefficients thus have general validity. Second, it is possible to obtain a number of important analytical results for the Spearman and Kendall correlation coefficient measures, whereas for the Pearson correlation no simple statements can be shown regardless of the distributions. However, there is a relation between the correlation measures and, to some extent, a result obtained for one correlation may be transferred to another one.

3.1. Spearman-uncorrelated pair of variables

One of the target error measures is the maximum deviation of the target and actual correlation matrix $\rho_{\max}$. Let us analyze the case of an uncorrelated target random vector characterized by the (unit) correlation matrix with *Spearman's correlation coefficients*. Clearly, the maximum correlation error $\rho_{\max}$ for uncorrelated variables equals one. Let us now focus on the lower bound of the error. The sum $\sum d_s^2$ in the numerator of Eq. (10) is an even number; see Eq. (12). On the other hand, the number $N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)/6$ may be either an even or an odd number, depending on N_{sim}. If $N_{\text{sim}} = 2+4l$, where l is a nonnegative integer, the quantity $N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)/6$ is an odd number. Otherwise, it is an even number leaving a chance that the numerator matches the denominator and $A_{i,j} = 0$. The errors ρ_{rms} and $\rho_{\max}$ match in the case of two random variables (see Eqs. (3) and (4)), and they are equal to the absolute difference of the correlations $|T_{i,j} - A_{i,j}|$. By substituting this into Eq. (10) we can, after some algebra, write

$$\rho_{\max} \frac{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)}{6} = \left| \left(1 - T_{i,j}\right) \frac{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)}{6} - \sum_{s=1}^{N_{\text{sim}}} d_s^2 \right|. \tag{18}$$

Since the target correlation coefficient $T_{i,j} = 0$, the above equation simplifies into

$$\rho_{\max} \frac{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)}{6} = \left| \frac{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)}{6} - \sum_{s=1}^{N_{\text{sim}}} d_s^2 \right|.$$

Two different minima are possible for the term on the right hand side of the equation depending on whether we subtract the even sum from an odd or an even number. The term can best be equal to 1 if $N_{\text{sim}} = 6 + 4l$; otherwise, it is equal to zero (recall the lower bound on the sum $\sum d_s^2$). From this we conclude that the optimal solution, or the lower bound on error $\rho_{\max}^{\text{Spear}}$ in Spearman correlation, is

$$\min \rho_{\max}^{\text{Spear}}(0) = \begin{cases} \dfrac{6}{N_{\text{sim}}\left(N_{\text{sim}}^2 - 1\right)} & \text{if } N_{\text{sim}} = 6 + 4l, \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

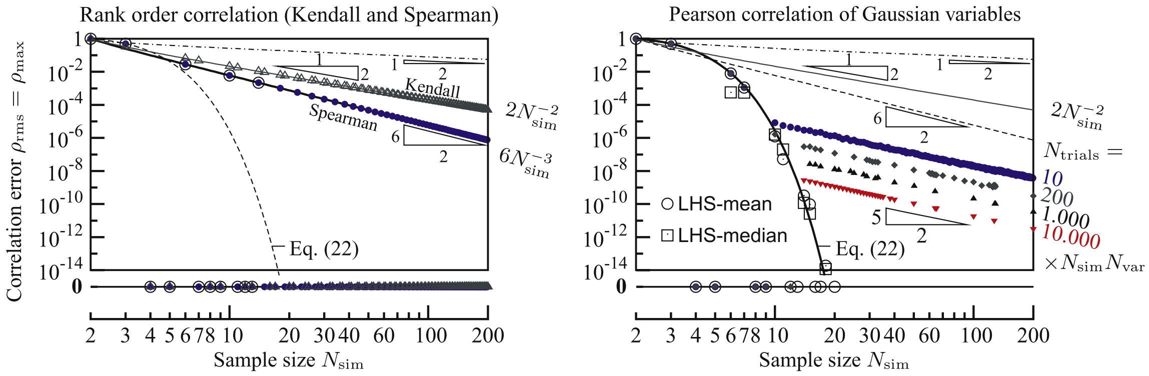

Fig. 1. Correlation error when the target correlation $T_{i,j} = 0$ (logarithmic graph with added zero ordinate). Solid symbols denote average algorithm performance. The uppermost dashed-dot line stands for the average correlation error arising from random ordering; see Eq. (45).

where l is a nonnegative integer. We now confirm this result by a different argument. Note that requesting zero Spearman correlation $A_{i,j}$ in Eq. (9) is equivalent to requesting that the sum of products of rank values equals

$$\sum_{s=1}^{N_{\rm sim}} \pi_{i,s}\,\pi_{j,s} = \frac{N_{\rm sim}\,(N_{\rm sim}+1)^2}{4}. \tag{20}$$

The left hand side is always an integer. If $N_{\rm sim} = 2 \pmod 4$, the right hand side is not an integer and the zero correlation cannot be achieved (modulo operation finds the remainder of the division of one number by another, written in parenthesis). For large $N_{\rm sim} = 2 \pmod 4$ we can approximate Eq. (19) and write the formula for correlation error as $6\,N_{\rm sim}^{-3}$. This formula is approximate, yet very accurate. In other words, the worst convergence of the best results is polynomial (a power law with an exponent of −3), and the associated error graph in a double logarithmic plot is a decreasing straight line of the same slope; see Fig. 1 (left).

3.2. Kendall-uncorrelated pair of variables

In analogy with Section 3.1, which is devoted to Spearman's correlation, we now analyze the smallest possible absolute values of Kendall's coefficient for a pair of variables. Kendall's tau can only attain zero when the numerators and denominators in Eq. (16) match: $4\,n_c = N_{\rm sim}\,(N_{\rm sim}-1)$ or $4\,n_d = N_{\rm sim}\,(N_{\rm sim}-1)$ (or equivalently when $n_c = n_d$). The number n_c $[n_d]$ can only take integer values and therefore the ratio $N_{\rm sim}\,(N_{\rm sim}-1)\,/4$ must be an integer to achieve zero correlation. This only happens when either $N_{\rm sim}$ or $(N_{\rm sim}-1)$ is divisible by four. Otherwise, the value of Kendall's correlation closest to zero is achieved for a unit difference between n_c and n_d, leading to correlations of $\pm 2/[N_{\rm sim}\,(N_{\rm sim}-1)]$. To conclude, the smallest correlation error of the proposed algorithm in the case of targeting *two Kendall-uncorrelated variables*:

$$\min \rho_{\rm max}^{\rm Kendall}(0) = \begin{cases} 0 & \text{if } N_{\rm sim} = 4l, \\ 0 & \text{if } N_{\rm sim} = 4l+1, \\ \dfrac{2}{N_{\rm sim}\,(N_{\rm sim}-1)} & \text{otherwise,} \end{cases} \tag{21}$$

where l is a nonnegative integer. This bound is presented in Fig. 1 (left).

3.3. Pearson-uncorrelated pair of variables

When any algorithm is used to reorder samples of Gaussian variables to obtain *Pearson-uncorrelated variables*, the situation is different. When $N_{\rm sim} = 4$ the sample set of a pair of Gaussian LHS-sampled vectors consists of the values $\{-b, -a, a, b\}$. If the first vector $\vec{\imath}$ remains unchanged, the second can be reordered into vector $\vec{p}$ or $\vec{q}$, both giving zero Pearson correlation with $\vec{\imath}$; see Table 1. Whenever $N_{\rm sim}$ is divisible by four, one can find these quaternaries consisting of two symmetrical pairs and solutions leading to zero Pearson correlation. Also, for each such sample size, one can also increase $N_{\rm sim}$ by one (add the zero sample value to both variables; see the very last row of Table 1). This preserves the zero correlation, and that is why we can conclude that whenever $N_{\rm sim}$ or $(N_{\rm sim}-1)$ is divisible by four, the zero target Pearson correlation can be matched exactly for any symmetrically distributed pair of samples of variables.

Table 1
Two pairs of Pearson-uncorrelated vectors.

$\vec{\imath}$	$\vec{p}$	$\vec{q}$	$\vec{\imath}\cdot\vec{p}$	$\vec{\imath}\cdot\vec{q}$
$-b$	$-a$	a	ab	ab
$-a$	b	$-b$	$-ab$	$-ab$
a	$-b$	b	$-ab$	$-ab$
b	a	$-a$	ab	ab
0	0	0	Sum = 0	Sum = 0

In the rest of the cases ($N_{\rm sim} = 2, 3; 6, 7; 10, 11; \ldots$), the lowest possible Pearson correlation must be estimated using a different argument. In particular, it is beneficial to estimate the distance from zero to the nearest attainable correlation. This is performed in the following subsection. Here, we conclude by writing the formula for the correlation error of the proposed algorithm in the case of targeting *two Pearson-uncorrelated variables*:

$$\min \rho_{\rm max}^{\rm Pears}(0) \begin{cases} = 0 & \text{if } N_{\rm sim} = 4l, \\ = 0 & \text{if } N_{\rm sim} = 4l+1, \\ \approx 1/\Gamma\,(N_{\rm sim}) & \text{otherwise,} \end{cases} \tag{22}$$

where l is a nonnegative integer and Γ is the Gamma function (the following equality $\Gamma\,(n) = (n-1)!$ holds for natural numbers, n). Note that whenever Pearson's correlation can achieve zero, Spearman's correlation can surely equal zero as well (compare Eqs. (19) and (22)). This can be easily explained: if the sample values are arranged so that they cancel each other out in the summation of products (recall the computation of Pearson's correlation, or better: covariance), the transformed ranks r must cancel each other out in Eq. (8), as well. The opposite does not hold when $N_{\rm sim} = 3 \pmod 4$. To conclude, when $N_{\rm sim} = 2 \pmod 4$ neither Spearman's nor Pearson's correlation can match zero exactly.

The quality of the predictions of the error can be assessed in Fig. 1. It can be seen that in the Spearman case, the predictions are exact (see on the left). In the Pearson case (right hand side), the accuracy of prediction by Eq. (22) depends on the choice of the sampling scheme such as LHS-mean or LHS-median; see [1]. In the numerical simulations, once $N_{\rm sim}$ exceeds a critical number of approximately 15, the predicted error is much lower than

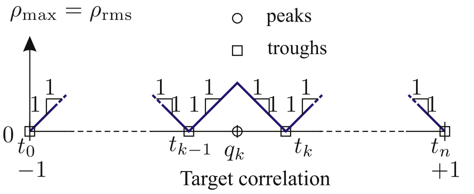

Fig. 2. The general shape of the lower bound on correlation errors for the whole spectrum of target correlations.

the actual results of the algorithm from Part I [1]. The average algorithm performance tends to follow a power law, i.e. a straight line with a decreasing slope of $\approx -\frac{5}{2}$. This sudden change of error rate must be attributed to the algorithm and will be analyzed in Part III [18]. The error rate of order $\mathcal{O}\left(N_{\mathrm{sim}}^{-5/2}\right)$ will be shown to be present also in multivariate settings. We have found that the vertical position of the straight line, and therefore also the critical sample size, N_{sim}, depends on the number of trials N_{trials} at a given temperature (a trial represents a random swap of two sample values; see Part I). The higher the number of trials, the better the result (lower correlation error). The number of trials recommended in Part I equals $N_{\mathrm{trials}} = 10 N_{\mathrm{sim}} N_{\mathrm{var}}$. This number should rather be proportional to $N_{\mathrm{sim}}!$, which would reflect the combinatorial nature of the problem. However, this would lead to explosion in computational demands with a large sample size. We have found that for a good balance between acceptable accuracy and computational expenses the above N_{trials} is a good choice.

A final note regarding Fig. 1 (right) is that the best solutions, represented by empty circles and boxes in the Pearson case, were obtained only for $N_{\mathrm{sim}} < 20$. To obtain exact results for sample sizes ranging from 15 to 20 we had to resort to a full permutation search. For greater N_{sim} the search becomes very expensive and also the accuracy (numbers of orders lower than 10^{-15}) starts to seriously depend on the accuracy of the evaluation of the inverse transformation of the standardized Gaussian distribution employed for sample value determination.

3.4. Generally correlated pair of variables

The above results regarding the lower bounds of correlation error with two Spearman or Kendall or Pearson-uncorrelated variables (Eqs. (19), (21), (22)) do not hold for a general value of the target correlation coefficient $T_{i,j}$. The problem is that, in a general case, the target value of the fraction in Spearman's or Kendall's correlation coefficient (Eqs. (10), (16)) is not equal to unity anymore. Similarly, given two vectors containing samples of random variables, only selected Pearson correlations can be achieved by permuting the mutual ranks of samples. From here on, we will call those *T*s that allow $\rho_{\max}(T) = 0$ '*troughs*' and denote them by t.

This paragraph is valid for all three studied correlation coefficients: Spearman, Kendall and Pearson. As explained above, some target correlations exist (the troughs, t) that can be fulfilled exactly by permuting mutual sample ranks. The number of troughs t_k, $k = 0, \ldots, n$, for a given sample size N_{sim}, is $n_n = (n+1)$. The first [last] troughs are $t_0 = -1$ [$t_n = +1$], irrespective of the number of simulations N_{sim}. Between any pair of consecutive troughs, t_k and t_{k+1}, the target correlation cannot be fulfilled exactly ($\rho_{\max} > 0$). It can be shown that in the middle of the distance between these pairs there are points in which the best achievable correlations have peaks. We call these points the *peaks* q_k: $q_k = \frac{1}{2}(t_{k-1} + t_k)$, $k = 1, \ldots, n$. This situation is depicted in Fig. 2. The error $\rho_{\max}$ is a linear function of the distance from its zero value in the troughs. The slope of the straight line is either ±1. That is why the best correlation error for any peak is equal to its distance to the nearest trough: $\rho_{\max}(q_k) = \frac{1}{2}(t_k - t_{k-1}) = q_k - t_{k-1}$. The profile of the lower bound on error ρ_{rms} is a piecewise linear function for the whole range of $T_{i,j} \in \langle -1; +1 \rangle$, alternately connecting troughs and peaks. It can be shown that the graph is symmetric with respect to the zero correlation $T_{i,j}$ (the solution of the best error for a given $T_{i,j}$ holds also for $-T_{i,j}$). The knowledge of the troughs t_k, $k = 0, \ldots, n$ for each sample size gives full information on the lower bound of the correlation norm $\rho_{\max}$.

3.4.1. Spearman correlation

For Spearman's correlation, an analysis of the best achievable correlation norm $\rho_{\mathrm{rms}} = \rho_{\max}$ can be easily carried out by analyzing Eq. (18). First, by exploiting the bounds on the sum $\sum d_s^2$ (Eq. (13)), we can see that the error ρ_{rms} can equal zero for $T_{i,j} = \pm 1$. The number of troughs (associated with the different possible results of the sum $\sum d_s^2$) is much smaller than the amount of $N_{\mathrm{sim}}!$ permutations (i.e., the number of different mutual orderings of two rank columns). In fact, n is much less than that and is $\mathcal{O}(N_{\mathrm{sim}}^3)$. This result can be obtained by analyzing all possible values of the sum $\sum d_s^2$. It can be shown that for every positive integer N_{sim} the sum $\sum d_s^2$ can take all even numbers from the interval given in Eq. (13). An exception occurs when $N_{\mathrm{sim}} = 3$, as for this the sum D cannot equal four and therefore there are only four troughs: −1; −0.5; 0.5; 1; see Fig. 3 (left).

For Spearman's correlation, there are $n_n = n + 1$ uniformly distributed troughs t_k, $k = 0, \ldots, n$ within $\langle -1; +1 \rangle$ associated with even numbers within the bounds from Eq. (13), where $n = N_{\mathrm{sim}}(N_{\mathrm{sim}}^2 - 1)/6$ is the number of positive even numbers from that interval except zero. It can be shown that the best correlation error also exhibits good symmetry. For a given permutation π_s of sample ranks (s) of the second variable, it can be shown that its correlation error from identity is equal to one minus the correlation error from reverse ranks ($r = N_{\mathrm{sim}} - s + 1$). This follows from the equality

$$\sum_{s=1}^{N_{\mathrm{sim}}} (s - \pi_s)^2 = u_N - \sum_{s=1}^{N_{\mathrm{sim}}} (r - \pi_s)^2 \tag{23}$$

where π_s is an arbitrary permutation of ranks $s = 1, \ldots, N_{\mathrm{sim}}$. The proof of this statement can be obtained by noting that $\sum (s - \pi_s)^2 + \sum (r - \pi_s)^2 = \cdots = 4 \sum s^2 - 2(N_{\mathrm{sim}} + 1) \sum s$, which is independent of the chosen permutation π_s. It therefore must be equal to the sum of the zero lower bound and the upper bound u_N (identical and reverse ranks). This can be used to prove the upper bound in Eq. (13). The consequence of the symmetry is that the problem of finding the best permutation of ranks for a given Spearman correlation coefficient $T_{i,j}$ collapses into finding a solution for the absolute value of $T_{i,j}$, and in the case of a negative target correlation, one must reverse the ranks.

The errors at the peaks are $6/\left(N_{\mathrm{sim}}^3 - N_{\mathrm{sim}}\right)$ again; compare with Eq. (19). The surface of the lower bounds on the correlation error $\min \rho_{\max}^{\mathrm{Spear}}(T_{i,j})$ is therefore fully characterized; see Fig. 3 (left and top). Numerical results from simulations with the proposed algorithm match this surface. Note that the peaks of correlation errors are polynomials in N_{sim}, with approximately the power of minus three, and therefore the graphs of peaks versus N_{sim} in a logarithmic plot again display what is approximately a straight line with a slope of minus three; see Fig. 1.

3.4.2. Kendall correlation

As for Kendall's sample correlation, the range $\langle -1; +1 \rangle$ is filled with equidistant attainable correlations t associated with all possible integer numerators in Eq. (16). In this respect, the situation is similar to Spearman correlation. The distance between any two adjacent correlations t_k and t_{k+1} is

$$\Delta t_{\mathrm{Kend}} = 4/(N_{\mathrm{sim}}^2 - N_{\mathrm{sim}}). \tag{24}$$

The number of attainable correlations (troughs) t_k, $k = 0, \ldots, n$ is $n_n = n + 1 = 2/\Delta t_{\mathrm{Kend}} + 1 = \binom{N_{\mathrm{sim}}}{2} + 1$.

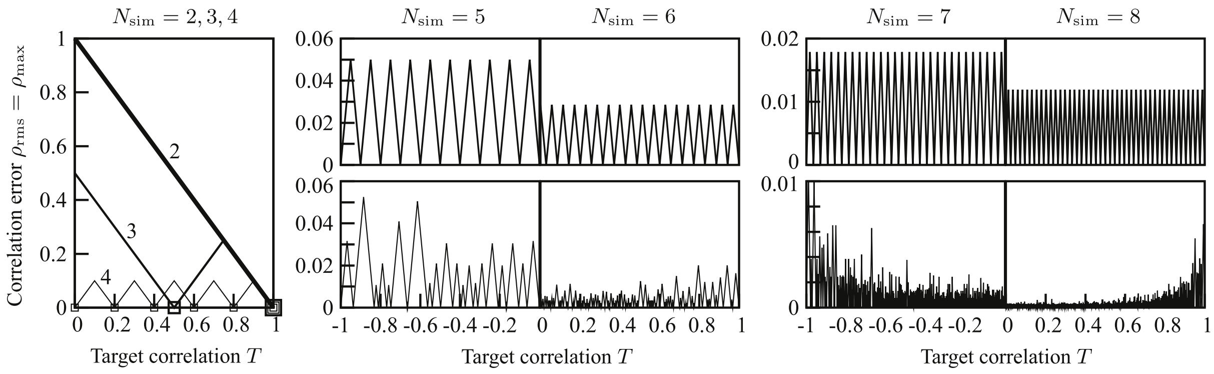

Fig. 3. Attainable correlations and minimum error for the full range of target correlations T and various sample sizes. Top: Spearman correlation. Bottom: Pearson correlation of Gaussian variables sampled via LHS-mean.

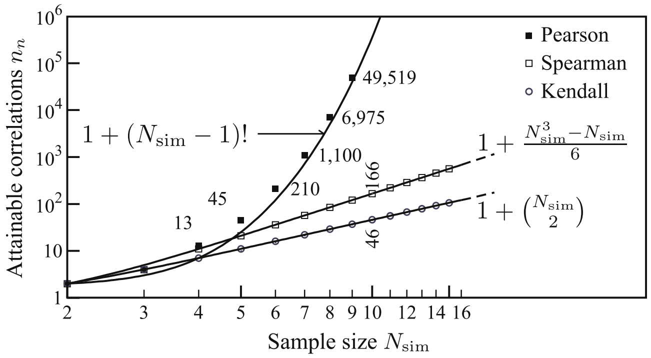

Fig. 4. Number of attainable correlations (troughs) t.

3.4.3. *Pearson correlation*

We now turn our attention to Pearson's correlation. The correlation estimation formulas are identical for Spearman and Pearson coefficients except that Spearman works with uniformly distributed integer values (ranks) while Pearson works with nonuniform real numbers. This difference suggests that the number of achievable correlations will be greater in the Pearson case. The peaks and troughs happen at different correlations T. The positions of troughs (and therefore also of peaks and errors at peaks) depend on the distribution of sampled variables and on the sampling scheme. For example, in the case of the compared sampling schemes LHS-median and LHS-mean, the numerically obtained trough distributions for the two Gaussian variables differ. The troughs are generally not uniformly distributed between -1 and $+1$ (see Fig. 3 (bottom)) and therefore the minimum errors $\rho_{\max}$ at peaks vary over the correlation range. Numerical simulations show that the greatest gaps between peaks and troughs occur in the vicinity of $T = -1$ and $+1$. This means that peaks are greater in the vicinity of these correlations than for $T_{i,j}$s close to zero; see Fig. 3 (bottom right).

Since the number of troughs (the number of different attainable correlations by rank permutations) is much greater in the Pearson case compared to the Spearman and Kendall cases, the spacing between them is less and therefore the peak errors are smaller as well. That is why, overall, the surface of lower bounds on correlation error is lower in the case of Pearson's correlation, compare Eqs. (19), (21) and (22). Fig. 4 compares the numbers of attainable correlations for Spearman and Kendall coefficients and with a Pearson coefficient obtained with Gaussian variables sampled via LHS (LHS-mean and LHS-median give identical numbers).

Fig. 1 and the related Eq. (22) shows that the worst errors at the peaks of Pearson pair correlation are well approximated by $1/\,(N_{\rm sim} - 1)!$. This number can be used to estimate the number of attainable correlations within the interval $\langle -1, 1\rangle$. If we consider equidistant spacing between troughs (which is not true exactly), the distance between the troughs must be twice the error at the peaks (see Fig. 2). The number of peaks must then be two over the distance, because the correlation range equals two. Then, the number of Pearson troughs must be approximately equal to one plus the inverse of the error at the peaks, i.e. $n_n = 1 + (N_{\rm sim} - 1)!$. The fact that for small $N_{\rm sim}$ this function does not agree with the numerically obtained numbers n_n in Fig. 4 can be explained by the nonuniform distribution of troughs in these cases. For greater $N_{\rm sim}$, however, the trend seems to be captured well.

We note that an issue to consider is the accuracy of the inverse transformation of the Gaussian cdf which is used to obtain the sample set. These values are used directly to compute Pearson's correlation and therefore the number of troughs is very sensitive to numerical accuracy when $N_{\rm sim}$ is large. This problem is not present in the case of Spearman and Kendall correlations.

4. Distribution of a random correlation

In this section, we study the frequency of the attainable Spearman, Pearson and Kendall correlations that appear in *random mutual ordering* of pairs of vectors representing random variables. It is a common practice to sample values from random vectors with independent marginals separately for each random variable, i.e. without employing any special technique for correlation control; see e.g. [19]. We begin with the analysis of a random correlation between two random variables (the three analyzed types of correlation coefficients are studied separately). Multivariate cases are then studied with the help of the defined norms of correlation matrices $\rho_{\rm rms}$ and $\rho_{\max}$.

4.1. *Spearman correlation of two random variables*

In the case of Spearman correlation, this task is equivalent to the study of the frequency of possible values of the sum D, Eq. (11). For each $N_{\rm sim}$ one can permute the rank orders against identity and compute the sum D. For $N_{\rm sim} \lesssim 14$, it is cheap to simply test all $N_{\rm sim}!$ permutations. The probability of any particular value of D (or a correlation t, a discrete random variable) is proportional to the number of permutations giving rise to this value. This probability is computed as $p_k = n_k/N_{\rm sim}!$, where $k = 0, \ldots, n$; here, $n = u_N/2$ and n_k is the number of occurrences of the kth value of the sum D_k (or the trough t_k). When $N_{\rm sim}$ is high, the probabilities can estimated as $p_k \doteq n_k/N_{\rm trials}$, where $N_{\rm trials}$ is the number of tested random permutations.

Let us note that the enumeration of probabilities for a few very large absolute correlations can also be performed analytically. First, the values $t = \pm 1$ corresponding to identical or reverse orders each have a probability of $p_0 = 1/N_{\rm sim}!$. The next possible value of

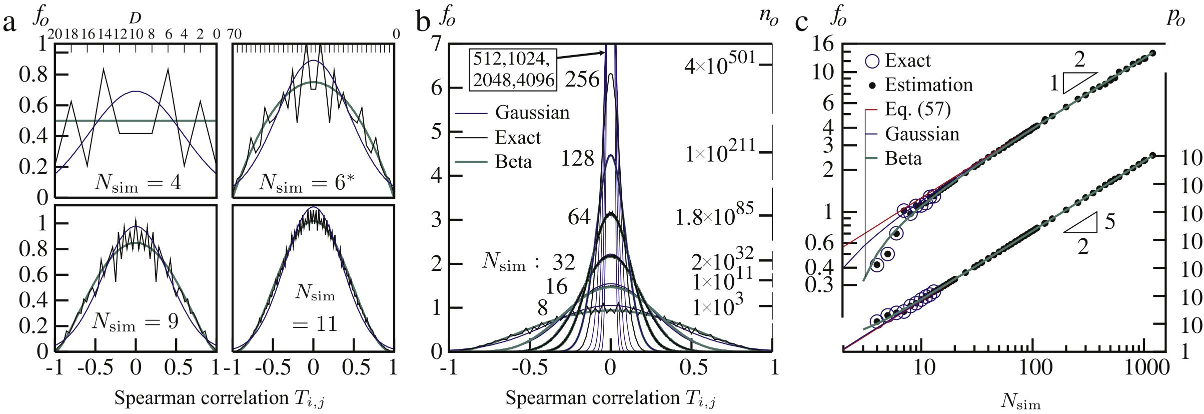

Fig. 5. Distribution of a random Spearman correlation. (a) and (b) exact histogram polygons compared with Gaussian and Beta distributions; (c) number of optimal solutions.

the sum $D_1 = 2$ corresponds to an interchange of two consecutive rank numbers which can be carried out for $N_{\mathrm{sim}} - 1$ different pairs. Thus, $p_1 = (N_{\mathrm{sim}} - 1)/N_{\mathrm{sim}}!$. Next to it, $D_2 = 4$, and by counting the number of ways of performing two interchanges of pairs of adjacent terms, the frequency to be divided by $N_{\mathrm{sim}}!$ is: $(N_{\mathrm{sim}} - 2)(N_{\mathrm{sim}} - 3)/2 = \binom{N_{\mathrm{sim}}}{2}$. For $D_3 = 6$ the frequency is $\binom{N_{\mathrm{sim}}-3}{3} + 2\binom{N_{\mathrm{sim}}-2}{1}$, for $D_4 = 8$ we have $\binom{N_{\mathrm{sim}}-4}{4} + 4\binom{N_{\mathrm{sim}}-3}{2} + \binom{N_{\mathrm{sim}}-2}{1}$, for $D_5 = 10$ we have $\binom{N_{\mathrm{sim}}-5}{5} + 6\binom{N_{\mathrm{sim}}-4}{3} + 2\binom{N_{\mathrm{sim}}-3}{2} + 2\binom{N_{\mathrm{sim}}-3}{1}$, etc. Expressions rapidly become very complicated. Next, we resort to another technique to quantify the probabilities p_k.

The range of D (and thus also the $\langle -1; 1\rangle$ range of t) gets filled very quickly as N_{sim} increases. For large N_{sim} it is easier to work with the distribution of correlations as with a continuous variable than with a discrete variable. The continuity correction can be made simply by assuming that the range of D (Eq. (13)) is equivalent to the range of correlations. The distance between any two consecutive Ds is two; therefore, the distance between adjacent correlations t_k and t_{k+1} is

$$\Delta t_{\mathrm{Spear}} = 4/u_N = 12/(N_{\mathrm{sim}}^3 - N_{\mathrm{sim}}). \tag{25}$$

The continuous probability density function (pdf) of correlations can be obtained from the discrete probability mass function as

$$f_E(t_k) = \frac{p_k}{\Delta t_{\mathrm{Spear}}} = p_k \frac{\left(N_{\mathrm{sim}}^3 - N_{\mathrm{sim}}\right)}{12}. \tag{26}$$

The numerically obtained histogram polygons are plotted, for selected N_{sim}, in Fig. 5a. As N_{sim} increases, the serrated profile of histograms gets smoothed from the tails towards the core range. In the same figure, we show a comparison with two continuous distributions, namely Beta and Gaussian distributions, which are discussed next.

The distribution of t must be symmetric around the zero mean value. The dispersion of D equals $N_{\mathrm{sim}}^2(N_{\mathrm{sim}}+1)^2(N_{\mathrm{sim}}-1)/36$, from which the standard deviation of a random Spearman correlation is

$$\sigma_t^{\mathrm{Spear}} = \frac{1}{\sqrt{N_{\mathrm{sim}} - 1}}, \tag{27}$$

a formula obtained by Student and incorporated in Pearson's [20] memoir, see [21].

Using exact simulations for N_{sim} up to 14 and using a statistical analysis of 10^9 simulations for $N_{\mathrm{sim}} \in \langle 14; 2000\rangle$, we have found that the pdf of random Spearman correlation can be nicely fitted by Beta distribution $t \sim B(\alpha, \beta, a, b)$ with the following pdf

$$f_B(t) = \frac{1}{(b-a)\,B(\alpha,\beta)} \left(\frac{t-a}{b-a}\right)^{\alpha-1} \left(\frac{b-t}{b-a}\right)^{\beta-1} \tag{28}$$

where the Beta function $B(\alpha, \beta) = \Gamma(\alpha)\,\Gamma(\beta)/\Gamma(\alpha+\beta)$ appears as a normalization constant to ensure that the total probability integrates to unity. The four parameters can be set using the following arguments. The location parameters (bounds) are the bounds of correlation: $a = -1$, $b = +1$. Since the distribution must be symmetric (see Eq. (23)), the two shape parameters must match: $\alpha = \beta$. Finally, by exploiting the result regarding the standard deviation (Eq. (27)), one finds that $\alpha = \frac{1}{2}(N_{\mathrm{sim}} - 2)$. The general formula for Beta distribution in Eq. (28) therefore collapses into one parameter form

$$f_B(t; N_{\mathrm{sim}}) = \frac{\Gamma(N_{\mathrm{sim}} - 2)}{\Gamma^2\left(\frac{1}{2}N_{\mathrm{sim}} - 1\right)} \frac{\left(1 - t^2\right)^{\frac{1}{2}N_{\mathrm{sim}}-2}}{2^{N_{\mathrm{sim}}-3}}. \tag{29}$$

This formula can be written using the Beta function $B(\cdot)$ as

$$\frac{\left(1-t^2\right)^{\frac{1}{2}N_{\mathrm{sim}}-2}}{B\left(\frac{1}{2}, \frac{1}{2}N_{\mathrm{sim}} - 1\right)},$$

which was previously obtained by Pitman [22] by noting that the first four moments of t are very close to the Pearson Type II curve. The Beta approximation is very good already for N_{sim} as small as 6; see Fig. 5a. For very large N_{sim} (hundreds or more), it is useful to use the result obtained by Hotteling and Pabst [21], who have proved that the asymptotic distribution ($N_{\mathrm{sim}} \to \infty$) of Spearman's correlation is Gaussian. Using the zero mean and standard deviation according to Eq. (27) we may write the Gaussian pdf as

$$f_G(t; N_{\mathrm{sim}}) = f_G(0) \exp\left(t^2 \frac{1 - N_{\mathrm{sim}}}{2}\right) \tag{30}$$

where $f_G(0)$ is the peak pdf of the Gaussian approximation

$$f_G(0) = \sqrt{\frac{N_{\mathrm{sim}} - 1}{2\pi}}. \tag{31}$$

For $N_{\mathrm{sim}} > 4$ the Beta distribution in Eq. (29) is unimodal, the mode being zero, and the peak pdf reads

$$f_B(0) = \frac{\Gamma\left(\frac{N_{\mathrm{sim}}-1}{2}\right)}{\sqrt{\pi}\;\Gamma\left(\frac{N_{\mathrm{sim}}}{2} - 1\right)}. \tag{32}$$

The motivation for the above-cited papers (or the papers cited in the section concerned with the random correlation distribution) on the distribution approximation of random Spearman correlation was mainly the significance testing of such correlation; see also [23–34] for the most important papers in this area. We, however, have the ambition to explore the probabilities of various correlations even for a multidimensional setting. In particular, we will focus on the probability of obtaining a (nearly) zero correlation in Section 5.1. The pdf at the mode will become very important there, because we will count the number of solutions yielding uncorrelatedness.

4.2. Pearson correlation of two random variables

As pointed out by Kowalski [35], the exact sampling distribution of the Pearson correlation coefficient for N_{sim} samples from a bivariate normal distribution has already been obtained by Fisher [36]. In the case of uncorrelated variables, the density of t becomes

$$f_B\left(t; N_{\text{sim}}\right) = \frac{\Gamma\left(\frac{1}{2}N_{\text{sim}} - \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}N_{\text{sim}} - 1\right)\sqrt{\pi}} \left(1 - t^2\right)^{\frac{1}{2}N_{\text{sim}}-2}, \tag{33}$$

which is identical to Eq. (29) written for the Spearman correlation (use the Legendre duplication formula for the Gamma function). From this we may conclude that the asymptotic distribution is again zero-mean Gaussian and the standard deviation from Eq. (27) holds also for the Pearson correlation coefficient

$$\sigma_t^{\text{Pears}} = \frac{1}{\sqrt{N_{\text{sim}} - 1}}. \tag{34}$$

4.3. Kendall correlation of two random variables

The sampling distribution of Kendall's tau has been found [6] to tend to a Gaussian distribution with zero mean and a standard deviation of

$$\sigma_t^{\text{Kendall}} = \sqrt{\frac{2\left(2N_{\text{sim}} + 5\right)}{9N_{\text{sim}}\left(N_{\text{sim}} - 1\right)}}. \tag{35}$$

As N_{sim} becomes large, the standard deviation becomes

$$\sigma_t^{\text{Kendall}} \approx \frac{2}{3}\frac{1}{\sqrt{N_{\text{sim}}}}. \tag{36}$$

The reason for the limiting standard deviation of the sampling Kendall correlation being about 2/3 of the Spearman correlation is a direct implication of the facts summarized at the end of Section 2.4. The distribution of Kendall correlation is very smooth and approaches a Gaussian distribution very rapidly compared to the Spearman coefficient. In fact, the Gaussian distribution is a very good approximation of the exact sampling distribution even for quite small N_{sim}.

4.4. Distribution of norms of a random correlation matrix

This section studies the distribution of the two particular norms defined in Part I as they occur when randomly permuting ranks of random variables. The target correlation matrix is the unit matrix $\boldsymbol{T} = \mathbf{I}$ (uncorrelated variables are desired) and therefore the error norms can be reformulated as norms of the actual correlation matrix $\boldsymbol{A}$ in the same way as was done in Section 2.1; see Eqs. (3) and (4). This is because when the correlation matrix entries $A_{i,j}$ are the results of random ordering of N_{sim} variable sample values for each of the N_{var} variables, we can understand the two matrix norms to be two statistics of $\boldsymbol{A}$ and work with them as with random variables.

It is immediately seen that the first norm, $\rho_{\max}$, is related to extremes (maxima) of N_c identically distributed variables (correlations) and that it will be more conservative than the second norm, ρ_{rms}, which is more of an average quantity. Unfortunately, the N_c variables are not independent. Why? Because given a random ordering of N_{var} samples (vectors of realizations; the first vector can be kept ordered for simplicity), only $N_{\text{var}} - 1$ correlation matrix entries are independent. The rest of the entries are dependent and this dependence influences the conditional distribution of the rest of the correlations. In addition, their distribution is not defined over the whole range of $\langle -1; 1\rangle$; it is defined over a subrange of it, because the correlation matrix $\boldsymbol{A}$ must be positive semidefinite (PSD).

From the geometry of correlation matrices it is known that PSD matrices form a solid body in the $[-1, 1]$ hypercube, where the coordinates are the correlations; see e.g. [37]. This body has its center at the origin (corresponding to the unit correlation matrix). If some correlations are given, the distribution of another entry in the matrix is generally a subrange of $\langle -1; 1\rangle$. However, our random given correlations have the majority of their pdf accumulated around the zero value and we expect that the conditional distribution of a new entry is not much different from the distribution of a random correlation (which is Gaussian in limit). Indeed, extensive simulation results show that many results obtained from the analysis of N_c independent and identically distributed (IID) correlations can be used.

In the remainder of this section, focused mainly on asymptotic properties, we work with the distribution of each correlation coefficient as with a Gaussian random variable with zero mean and variance σ_t^2. We know that σ_t depends solely on the sample size N_{sim} and the dependence for Spearman, Kendall and Pearson correlations is given by Eqs. (27), (35) and Eq. (34) respectively. As will become clear below, the asymptotic properties of errors are combinations of independent effects of the (i) sample size (through σ_t) and the (ii) number of variables N_{var}. The following subsections present results valid for Pearson and Spearman correlations, i.e. correlations that share the same formula for σ_t (compare Eqs. (27) and (34)). Derivation of the results for Kendall correlation can be easily obtained analogically by performing the same steps and employing Eq. (35) instead.

4.5. Distribution of ρ_{rms} for random ordering

The root mean square error ρ_{rms} (Eq. (4)) is very similar to the following function of a vector of N_c IID standard Gaussian variables

$$Z = \sqrt{\sum_1^{N_c} X_i^2}, \tag{37}$$

which is known to follow the χ (or Chi) distribution with the mean value μ_Z and variance σ_Z^2 given by

$$\mu_Z = \sqrt{2}\,\frac{\Gamma\left(\frac{N_c+1}{2}\right)}{\Gamma\left(\frac{N_c}{2}\right)}, \qquad \sigma_Z^2 = N_c - \mu_Z^2. \tag{38}$$

In fact, we analyze a linear transformation of Z: statistic $\rho_{\text{rms}} = aZ$, where $a = \sigma_t/\sqrt{N_c}$. This is because, at the limit, the random correlation $A_{i,j}$ follows a Gaussian distribution with zero mean and standard deviation σ_t. It is known that the mean value of ρ_{rms} is simply $\mu_{\text{rms}} = a\mu_Z$ and variance $\sigma_{\text{rms}}^2 = a^2\sigma_Z^2$. That is why the mean value of ρ_{rms} is

$$\mu_{\text{rms}} = \sigma_t \underbrace{\sqrt{\frac{2}{N_c}}\,\frac{\Gamma\left(\frac{N_c+1}{2}\right)}{\Gamma\left(\frac{N_c}{2}\right)}}_{r_\mu(N_c)} = \sigma_t \cdot r_\mu\left(N_c\right) \tag{39}$$

and the standard deviation of ρ_{rms} reads

$$\sigma_{\text{rms}} = \sigma_t\sqrt{1 - \frac{\mu_Z^2}{N_c}} = \sigma_t\,\underbrace{\sqrt{1 - r_\mu^2\left(N_c\right)}}_{r_\sigma(N_c)}. \tag{40}$$

The probability density function of ρ_{rms} for any N_c and input standard deviation σ_t of a random Gaussian correlation is the scaled χ distribution

$$f_{\text{rms}}\left(x; \sigma_t, N_c\right) = \frac{2^{\left(1-\frac{N_c}{2}\right)}}{\Gamma\left(\frac{N_c}{2}\right)}\left(\frac{x\sqrt{N_c}}{\sigma_t}\right)^{N_c}\frac{1}{x}\exp\left(-\frac{x^2N_c}{2\sigma_t^2}\right). \tag{41}$$

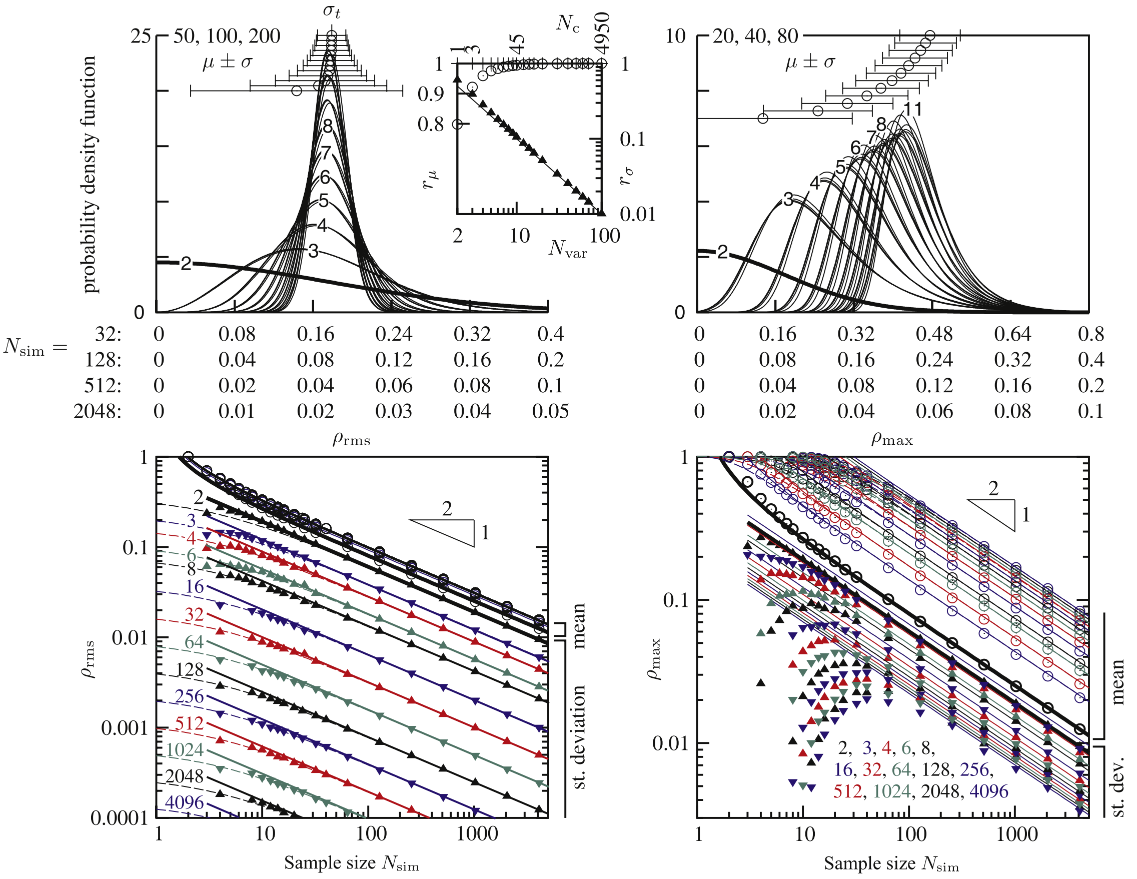

Fig. 6. Displays correlation norms ρ_{rms} (left column) and ρ_{max} (right column) when N_{var} arrays with N_{sim} values each permute randomly. Top row: stroboscopic evolution of the theoretical distributions for $N_{var} = 2, 11(1)$ and various sample sizes N_{sim}. Bottom row: comparison of numerically estimated mean values (empty circles) and standard deviations (solid triangles) with analytical formulas (see text).

It is interesting to see that the mean value (Eq. (39)) of the matrix norm ρ_{rms} is independently influenced by N_{sim} and by N_{var} (related to N_c through Eq. (2)). However, the mean value of the error, in fact, almost does not depend on $N_{var}[N_c]$ at all because r_μ quickly converges to unity; see the inset of Fig. 6 (top middle). On the contrary, the standard deviation is nearly exactly inversely proportional to the number of variables N_{var}, because $r_\sigma\ (N_c) \to {N_{var}}^{-1}$; see Eq. (40) and the inset of Fig. 6 (top middle). In other words, it might be surprising that, for a given sample size N_{sim}, adding more random variables does *not* increase the average error ρ_{rms}, and the standard deviation *decreases*:

$$\lim_{N_{sim}\to\infty} \mu_{rms} = N_{sim}^{-1/2} \tag{42}$$

$$\lim_{N_{sim}\to\infty} \sigma_{rms} = N_{sim}^{-1/2}\ {N_{var}}^{-1}. \tag{43}$$

Note that in the case of Kendall correlation coefficient, the right hand side of Eqs. (42) must be multiplied by $\frac{2}{3}$, see Eq. (36). Numerical simulations show that for an extremely small number of simulations ($N_{sim} < 10$) the standard deviation is slightly overestimated. Therefore, we recommend slightly changing Eq. (40) by replacing σ_t with $(N_{sim} + 3)^{-1}$; see the dashed lines in Fig. 6 (bottom left). Again, for Kendall correlation, one has to accordingly adjust Eq. (35) for σ_t.

Fig. 6 (top left) presents the distributions of ρ_{rms} for $N_{var} = 2, 11(1)$. It can be seen that the (properly scaled) distributions almost perfectly collapse into one for a given number of N_{var} and various sample sizes N_{sim}. The stabilization of the mean μ_{rms} is already visible for this small N_{var} value. The convergence of σ_{rms} to a power law can be seen by comparing it with the triangles in Fig. 6 (bottom left).

4.6. Distribution of ρ_{max} for random ordering

The first norm (Eq. (3)) represents the maximum of N_c absolute values of random correlations (entries in the upper triangle of the correlation matrix). We first derive the limit distribution of the absolute value of random correlation and then find the limiting distribution of the maxima.

The limiting pdf f_e of each correlation $A_{i,j}$ is Gaussian (f_G) with zero mean and standard deviation σ_t. Its absolute value $|A_{i,j}|$ must have a double density $g_{abs} = 2\ f_e$ over the half interval $\langle 0;\ 1\rangle$ (zero left bound). Both the (elemental) pdf (see curve '2' in Fig. 6 (top)) and the cumulative distribution function (cdf) can be expressed using the standard Gaussian pdf ϕ and cdf Φ as

$$g_{abs}\ (x) = \frac{2}{\sigma_t}\phi\left(\frac{x}{\sigma_t}\right), \qquad G_{abs}\ (x) = 2\Phi\left(\frac{x}{\sigma_t}\right) - 1 \tag{44}$$

and the mean value and standard deviation are

$$\mu_{abs} = \sigma_t\sqrt{\frac{2}{\pi}} \approx 0.8\ \sigma_t, \qquad \sigma_{abs} = \sigma_t\sqrt{\frac{\pi - 2}{\pi}} \approx 0.6\ \sigma_t. \tag{45}$$

It might be interesting to show that for two random variables $N_{var} = 2, N_c = 1$ the two norms match: $\rho_{rms} = \rho_{max}$. Indeed, the Half-Normal distribution is identical to the χ distribution with one degree of freedom: $f_{rms}\ (x;\ \sigma_t,\ 1) = g_{abs}\ (x)$ (compare Eqs. (41) and (44)). Inevitably in such a case, their moments also match: $\mu_{rms} = \mu_{abs}$ and $\sigma_{rms} = \sigma_{abs}$.

If we, again, consider the N_c entries of the correlation matrix to be a vector of IID variables, ρ_{max} is a random variable defined as the maximum of these IID variables; see Eq. (3). The cdf and pdf of the maxima of N_c IIDs read

$$F_{\max}(x) = H_n(x) = G_{\mathrm{abs}}^{N_c}(x) \tag{46}$$

$$f_{\max}(x) = \frac{\mathrm{d}F_{\max}(x)}{\mathrm{d}x} = N_c \; G_{\mathrm{abs}}^{N_c-1}(x) \; g_{\mathrm{abs}}(x) \,. \tag{47}$$

What is the limiting form of this distribution for large N_c? In order to avoid degeneration of the limit distribution we look for the linear transformation $Y = a_n + b_n x$, where a_n and b_n are constants depending on $n = N_c$ in such a away that the limit distribution $\lim_{n\to\infty} H_n(a_n + b_n x) = \lim_{n\to\infty} G_{\mathrm{abs}}^{N_c}(a_n + b_n x) = H(x)$ becomes non-degenerated. The extremal types theorem from the extreme value theory states an important result for a random variable $M_n = \max\{X_1, X_2, \ldots, X_n\}$ where $X_1, X_2, \ldots$ is a sequence of IID random variables. If two sequences of real numbers, a_n and b_n, exist so that $b_n > 0$, and $\lim_{n\to\infty} P[(M_n - a_n)/b_n \le x] = H(x)$ where H is a nondegenerate distribution function, then H must be one of the three feasible limit distributions for maxima, namely Gumbel, Fréchet or Weibull distribution (see [38] [39]; previous versions were stated by Fisher and Tippett in 1928 [40] and Fréchet in 1927 [41]). It can be easily checked that our distribution G_{abs} belongs to the Gumbel domain of attraction for maxima (see Eq. 3.18 of [42]) and therefore H is the standard Gumbel (max) distribution. The normalizing constants are the mode $m = a_n$ and the scale parameter $\beta = b_n$. The inverse of our elemental distribution is $F^{-1}(p) = \sigma_t \Phi^{-1}\left(\frac{p+1}{2}\right)$. The normalizing constants can be chosen as (see page 101 of [42]): $a_n = F^{-1}\left(1 - \frac{1}{n}\right)$ and $b_n = F^{-1}\left(1 - \frac{1}{ne}\right) - a_n$. In our case $a_n = \sigma_t \Phi^{-1}\left(1 - \frac{1}{2n}\right)$ and $b_n = \sigma_t \Phi^{-1}\left(1 - \frac{1}{2ne}\right) - a_n$. We now exploit the similarity of the two formulas with those written for the extremes of the standard Gaussian distribution: the differences are that (i) n is multiplied by two in our case and (ii) we work with the non-unit standard deviation σ_t. Therefore, we adapt the well known result for the standard Gaussian case (see e.g. Theorem 1.5.3. in [43]) so that

$$\beta_0 = \frac{1}{\sqrt{2w}}, \qquad m_0 = \sqrt{2w} - \frac{\ln(w) + \ln(4\pi)}{\sqrt{8w}}, \tag{48}$$

where $w = \ln(2N_c) = \ln\left({N_{\mathrm{var}}}^2 - N_{\mathrm{var}}\right)$. The corresponding mean value and standard deviation of the Gumbel approximation are also dependent solely on the number of variables (correlations) N_c:

$$\mu_0 = m_0 + \gamma \beta_0, \qquad \sigma_0 = \beta_0 \frac{\pi}{\sqrt{6}} \tag{49}$$

where $\gamma \doteq 0.577216$ is the Euler–Mascheroni constant. The solution for Gaussian elemental distribution with non-unit variance σ_t^2 is obtained simply by multiplying the moments and parameters with σ_t:

$$\mu_{\max} = \mu_0 \, \sigma_t, \qquad \sigma_{\max} = \sigma_0 \, \sigma_t \tag{50}$$

$$m_{\max} = m_0 \, \sigma_t, \qquad \beta_{\max} = \beta_0 \, \sigma_t. \tag{51}$$

Therefore, both the mean value and standard deviation are asymptotically inversely proportional to the square root of N_{sim} (because this is also true for σ_t; see Eqs. (27), (34), (35)). Similarly to ρ_{rms}, these formulas for the asymptotic mean and the standard deviation of $\rho_{\max}$ feature the independent effect of sample size N_{sim} and number of variables N_{var}. Unfortunately, for small N_{sim} the Gaussian standard deviation σ_t is relatively high and $\mu_{\max}$ exceeds unity whenever $N_{\mathrm{sim}} \approx {\sigma_t}^{-2} < \mu_0^2$. This happens more often for large N_c [or N_{var} respectively]. However, the real maximum correlation error $\rho_{\max}$ cannot exceed one. For this reason we propose a simple correction that 'bends' the straight mean curve at the turning point $N_{\mathrm{sim}} = \mu_0^2$ (see Fig. 6 (bottom right)) while retaining the asymptotic behavior for large N_{sim}:

$$\mu_{\max}^{\star} = \left(\frac{\mu_0^2}{\mu_0^2 + {\sigma_t}^{-2}} + {\sigma_t}^{-2} {\mu_0}^{-2} \right)^{-1/2} . \tag{52}$$

Also, numerical simulations show that the standard deviation is overestimated for small N_{sim} and large N_{var}: it must suddenly become zero when the mean approaches unity; see Fig. 6 (bottom right). However, we do not propose any correction for the standard deviation $\sigma_{\max}$.

Finally, the cdf and pdf of $\rho_{\max}$ can be approximated by the Gumbel distribution as

$$\lim_{N_c\to\infty} F_{\max}(x) = \exp\left[-\exp(-y)\right], \qquad y = \frac{x - m_{\max}}{\beta_{\max}}$$

$$\lim_{N_c\to\infty} f_{\max}(x) = \frac{1}{\beta_{\max}} \exp\left[y - \exp(-y)\right] \tag{53}$$

where the parameters are found by inverting Eq. (49):

$$\beta_{\max} = \frac{\sigma_{\max}\sqrt{6}}{\pi}, \qquad m_{\max} = \mu_{\max}^{\star} - \gamma \beta_{\max}. \tag{54}$$

Note that the Gumbel limit distribution is attained for very large N_{var}. However, simulation results show that the Gumbel approximation is reasonably good even for moderate N_{var}. For very small N_{var} the distribution shape differs from the Gumbel one and therefore the illustrative stroboscopic evolution of distribution for $N_{\mathrm{var}} = 2, 11(1)$ in Fig. 6 (top right) is not perfectly legitimate; it is displayed for the purpose of comparison with the evolution for ρ_{rms} in the same figure, top left. It should also be noted that for very small N_{sim} the Gaussian approximation with a standard deviation of σ_t to $A_{i,j}$ is inaccurate because the real frequency polygons of a random correlation have a serrated profile; see e.g. Fig. 5 (left).

4.7. Ratio of average errors $\rho_{\max}$ and ρ_{rms} for random ordering

The mean error μ_{rms} is asymptotically independent of the dimension of random vector N_{var}; see Eq. (42). The norm $\rho_{\max}$ is more conservative and its mean error increases with N_{var} (Eqs. (48)–(50)). Let us now study the asymptotic ratio between the two mean values. Both norms are $\propto N_{\mathrm{sim}}^{-1/2}$ and therefore their ratio will only be a function of the vector dimension N_{var}:

$$\begin{aligned} \nu_{\mu}(N_{\mathrm{var}}) &= \frac{\lim\limits_{N_{\mathrm{sim}}\to\infty} \mu_{\max}}{\lim\limits_{N_{\mathrm{sim}}\to\infty} \mu_{\mathrm{rms}}} = \frac{\mu_0 \, \sigma_t}{\sigma_t} = \mu_0 \\ &= \frac{4w + 2\gamma - \ln(w) - \ln(4\pi)}{\sqrt{8w}} \end{aligned} \tag{55}$$

where the meanings of w and γ remain as in Eqs. (48) and (49): $w = \ln(2N_c) = \ln\left({N_{\mathrm{var}}}^2 - N_{\mathrm{var}}\right)$; $\gamma \doteq 0.577216$ is the Euler–Mascheroni constant.

This ratio, obtained for random sample ordering, will be very important in the analysis of the algorithm's performance. We will show that the ratio between average errors remains *unchanged* even when the algorithm proposed in part I is employed and the convergence of both errors to zero is much faster (because the sample is not random there); see part III.

4.8. Comments

The target of any stochastic optimization algorithm for correlation control may be understood in this way: we want to move the mean values of both ρ_{rms} and $\rho_{\max}$ toward zero (i.e. toward the left in the two Fig. 6 (top)) and, at the same time, decrease the variability of those two error measures to zero (i.e. narrow the distributions in the same figure). If any algorithm for correlation control starts with a random vector order, remembers this state and eventually accepts only improvements, then the errors for which the distributions have been found in this section constitute the *upper bounds* of algorithm error.

The next section is focused on an analysis of the lower bound of ρ_{rms} and $\rho_{\max}$ error in a multivariate setting, i.e. on the best possible performance associated with the left bound of error distributions. It can be shown that for selected N_{sim} and N_{var} both the average error and the variance of errors can even be reduced to zero (we recall Section 3.1). This can be achieved by ordering the ranks (permutations) of sample values to yield *optimal solutions*.

5. Optimal solutions yielding perfect uncorrelatedness

The target of this section is to show that optimal orderings exist between vectors representing ranks of samples in the sense of either Pearson, Spearman or Kendall correlations. We also aim to count these optimal orderings. In the Pearson case by the term 'optimal orderings' we mean that all pairwise correlations are zero and when Spearman or Kendall correlation is considered we mean optimal orderings leading to the lower bound in the sense of Eqs. (19) or (21).

It should be noted that the primary task of an analyst can be to simulate samples from independent distributions. This is very often checked by estimating correlations and requesting uncorrelatedness. It is well known that zero correlation between two variables does not imply their mutual independence. Examples are easily constructed in which x is a function of y and yet the two variables are uncorrelated.

Before we proceed with the presentation of our results, we would like to point out that the problem of finding optimal solutions in the form of perfectly mutually orthogonal Latin Hypercube (LH) designs have been studied by many authors. The most important works are mentioned below.

Owen [44] and Tang [45] showed how orthogonal arrays can be used to generate LH designs with better properties than those of the original LHS method published by McKay et al. [46]. An algorithm that utilizes orthogonal arrays was presented in [45] with a simple algorithm to construct Orthogonal Latin Hypercubes. Later, Ye [47] presented a construction algorithm for a class of orthogonal LHs for $N_{\mathrm{sim}} = 2^m$ or $N_{\mathrm{sim}} = 2^m + 1$, $m > 2$ (the latter is constructed from the former by adding zeros as an additional simulation — done in the same way in Section 3.3). Using his algorithm he is able to find a solution for $N_{\mathrm{var}} = 2m$ variables (columns). Butler [48] showed how to construct LH designs in which terms in a class of trigonometric regression models are orthogonal to one another. Quite recently, Steinberg and Lin [49] discussed the weak points of Ye's approach. They have described a construction method for orthogonal Latin Hypercube designs based on a combination of two ideas co-authored by the two authors, the first being developed by Beattie and Lin [50] and the second one by Bursztyn and Steinberg [51]. The first idea is that a certain class of orthogonal LH-designs can be constructed as rotations of $2^{N_{\mathrm{var}}}$ factorial designs. The second idea is that rotations can be applied to groups of factors (variables), thereby greatly increasing the number of N_{var} in the resulting design. Using their method they are able to construct orthogonal LH-designs with $N_{\mathrm{sim}} = 2^{N_{\mathrm{var}}}$, where $N_{\mathrm{var}} = 2^m$. The number of possible N_{var} is almost as large as N_{sim}. However, the severe sample size constraint (N_{sim} must be powers of two) is the primary limitation of their design.

5.1. Spearman-uncorrelatedness

5.1.1. Pair of random variables

Table 2
Number of Spearman optimal solutions in the sense of Eq. (19).

N_{var}: N_{sim}	Two $n_o \equiv n_{o,2}$	Three $n_{o,3}$	Four $n_{o,4}$
4	2	0	0
5	6	0	0
6*	(2×)29	(See Table 3)	0
7	184	768	0
8	936	11,984	4,752
9	6,688	483,924	1099,352
10*	(2×)49,062	(See Table 3)	
11	420,480	942,553,720	
12	4298,664		
13	44,405,142		

Table 3
Number of Spearman optimal solutions $n_{o,3}$.

Accept corr.	Codes for variable p		
between p q	c_p^+	c_p^-	c_p^+ and c_p^-
Sample size $N_{\mathrm{sim}} = 6$:			
$\rho_{\max}^+$	0	0	48
$\rho_{\max}^-$	24	24	48
Both (sum)	24	24	96
Sample size $N_{\mathrm{sim}} = 10$:			
$\rho_{\max}^+$	15,296,207	15,296,207	60,806,010
$\rho_{\max}^-$	15,106,798	15,106,798	60,806,010
Both (sum)	30,403,005	30,403,005	121,612,020

It is interesting to study the number of optimal solutions to the combinatorial optimization problem, i.e. solutions leading to a Spearman correlation of zero or the one given in Eq. (19). The number of optimal solutions will be called n_o from here on. The subscript o stands for optimal and the corresponding sum D_o equals either $u_N/2 - 1$ when $N_{\mathrm{sim}} = 6 + 4l$ (resp. $u_N/2 + 1$ because there are two symmetric correlations), or $u_N/2$ otherwise. The number n_o grows *very* fast with N_{sim}; see Table 2. By plotting the probability p_o of receiving the optimal solution from Eq. (19) against N_{sim} in a double logarithmic plot, one can note that the graph looks like a straight line; see Fig. 5c. The asymptotic slope of this line can be deduced from the fact that the $p_o/\Delta t$ converge to the peak of Beta pdf which converges to the peak of Gaussian pdf approximating the empirical $f_E(t)$:

$$f_o = \frac{p_o}{\Delta t_{\mathrm{Spear}}} = \frac{n_{o,\mathrm{Spear}}}{N_{\mathrm{sim}}!}\,\frac{1}{\Delta t_{\mathrm{Spear}}} \approx f_B(0) \to f_G(0)\,. \tag{56}$$

Note that for very large N_{sim} we may say that $f_G(0) \propto N_{\mathrm{sim}}^{1/2}$ and therefore both the density and frequency of the optimal solutions may approximated by power laws

$$f_o \approx \frac{1}{\sqrt{2\pi}} N_{\mathrm{sim}}^{1/2}, \qquad p_o \approx \frac{12}{\sqrt{2\pi}} N_{\mathrm{sim}}^{-5/2} \tag{57}$$

and the asymptotic number of optimal solutions is

$$n_{o,\mathrm{Spear}} \approx \frac{12}{\sqrt{2\pi}}\,\frac{N_{\mathrm{sim}}!}{N_{\mathrm{sim}}^{5/2}} = \frac{12}{\sqrt{2\pi N_{\mathrm{sim}}}}\,\frac{\Gamma(N_{\mathrm{sim}})}{N_{\mathrm{sim}}}. \tag{58}$$

Using Stirling's formula we may write

$$n_{o,\mathrm{Spear}} \approx 12\,\frac{N_{\mathrm{sim}}^{N_{\mathrm{sim}}-2}}{\exp(N_{\mathrm{sim}})}, \tag{59}$$

which highlights the rapid increase of n_o with N_{sim}. The agreement with numerically obtained values of the peak densities f_o and relative frequencies p_o of optimal solutions can be assessed in Fig. 5c. Eq. (58) was used to compute the total number of optimal solutions for various N_{sim} that are highlighted in the right hand side of Fig. 5b. Eq. (58) quantifies the above statement that the number

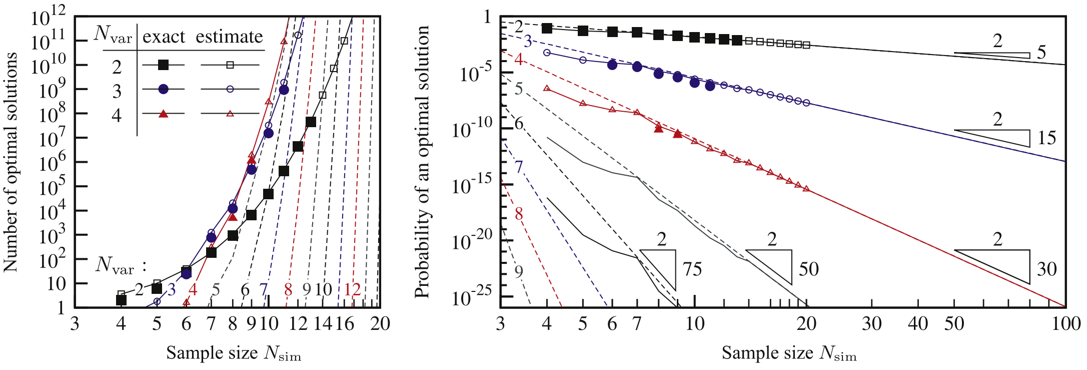

Fig. 7. Spearman optimal solutions for N_{var} random variables. Left: numbers of optimal solutions. Right: probabilities of optimal solutions. Dashed lines represent Eq. (64).

of optimal solutions (or simply zero correlations) grows very fast with N_{sim}. Not only that, the number of near optimal solutions also grows very fast. This can be seen from Fig. 5b, where the histograms of Spearman correlations become very narrow with high N_{sim}.

5.1.2. Three and more random variables

In the preceding section we have found that the number of optimal solutions grows fast with the sample size for a *pair* of random variables represented by two vectors of values. An important question is how many optimal solutions can be found when the number of random variables increases.

The number of all different mutual orderings for a pair of random variables is $N_{\mathrm{sim}}!$, because we keep the vector $\vec{\imath}$ representing the first random variable ordered (identity permutation) and permute the second vector, say $\vec{p}$. When the number of random vectors is N_{var}, the number of all different possible mutual orderings is

$$(N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1}. \tag{60}$$

The number given by Eq. (60) represents *all equiprobable possible correlation matrices* (related to all possible mutual orderings of samples) in the case when sample ordering is left random (as proposed e.g. in [19]).

Every permutation of a vector of N_{sim} values can be numbered by a code $c \in \langle 0; N_{\mathrm{sim}}!)$ that maps e.g. lexicographically ordered permutations. When $c = 0$ the permutation is ordered (identity permutation). The set of all $n_o \equiv n_{o,2}$ optimal solutions for a pair of vectors $\{\vec{\imath}, \vec{p}\}$ is therefore associated with a set of n_o pairs of codes $\{c_i, c_p\}$, where $c_i \equiv 0$ always and c_p are permutation numbers for variable p. Note that every $c_p > c_i$. We denote the set of all n_o codes c_p by Π_o.

We are now interested in finding all $n_{o,3}$ different optimal solutions for three random variables, i.e., all triples of permutation codes $\{c_i, c_p, c_q\}$ so that the correlations between all three pairs of $\{\vec{\imath}, \vec{p}, \vec{q}\}$ equal zero. The first vector $\vec{\imath}$ remains ordered ($c_i = 0$) and therefore uncorrelatedness between $\vec{\imath}$ and $\vec{p}$ is achieved by selecting the second codes, c_p, from Π_o. Similarly though, the third codes, c_q, are selected from Π_o. For each permutation code $c_p \in \Pi_o$ we seek optimal codes $c_q \in \Pi_o$ so that the Spearman correlation between $\vec{p}$ and $\vec{q}$ is zero. We also request that $c_p > c_q$ in order to avoid counting two mutual permutations that are obtained merely by interchanges of vectors $\vec{p}$ and $\vec{q}$. There can be, at most, n_o^2 permutations of $\vec{p}$ and $\vec{q}$ and this number decreases to $n_o(n_o+1)/2$ by removing the interchanges of $\vec{p}$ and $\vec{q}$. We can also exclude situations when $c_p = c_q$ (identical ordering) and also n_o situations of reverse orderings (which we know are present in Π_o). In any case, the estimate of the upper bound on $n_{o,3}$ remains proportional to $n_o^2/2$, which indicates that the number of optimal triples might be greater than the number of optimal pairs.

A similar procedure can be adopted when numerically seeking the number of optimal solutions for four random variables $n_{o,4}$. Every triple of permutation codes $\{c_i, c_p, c_q\}$ from the preceding paragraph can be tested against n_o codes $c_r \in \Pi_o$ for the fourth variable $\vec{r}$ — to obtain the uncorrelated pairs $\{\vec{p}, \vec{r}\}$ and $\{\vec{q}, \vec{r}\}$. Again, we request that every $c_r > c_q$, which is strictly greater than c_p. Clearly, for a high enough sample size, N_{sim}, the number of optimal solutions must grow with the number of random variables, N_{var}. On the other hand, when N_{sim} is very small adding too many variables must result in the loss of some optimal solutions otherwise obtained for lower N_{var}.

We used this procedure to numerically determine the numbers of the optimal solutions for small N_{sim} and N_{var}. The results are summarized in Table 2 and plotted in Fig. 7 (left). The numbers given in the table correspond to N_{var} orthogonal vectors that are sorted from the lexicographically smallest $\vec{\imath}$ through $\vec{p}, \vec{q}, \vec{r}, \ldots$ to the largest one. The total number of pairwise orthogonal vectors with the first vector $\vec{\imath}$ sorted would be obtained by also counting their interchanges, i.e. multiplying the numbers by

$$(N_{\mathrm{var}} - 1)!. \tag{61}$$

When $N_{\mathrm{sim}} = 6$ or 10, the optimum solution for a pair of variables is not unique (zero) and there are always two nonzero optima: $\rho^+_{\max} = +6/\left(N^3_{\mathrm{sim}} - N_{\mathrm{sim}}\right)$ and $\rho^-_{\max} = -6/\left(N^3_{\mathrm{sim}} - N_{\mathrm{sim}}\right)$ (see Eq. (19)). The two permutations are associated with the codes c^+_p and c^-_p. Therefore, we recognize the numbers n^+_o and n^-_o in these cases. The number of optimal triples then depends on whether we use c^+_p or c^-_p or both for the second variable p and also on whether we accept the positive correlation error or the negative error or both; see Table 3. For higher N_{sim} an exhaustive search is not feasible. Therefore, in order to estimate the numbers of optimal orderings and their probabilities another approach must be adopted.

In order to estimate the number of optimal solutions $n_{o,N_{\mathrm{var}}}$ it is useful to estimate the probability that one will hit an optimal solution (unit correlation matrix) when permuting N_{var} vectors: $p_{o,N_{\mathrm{var}}}$. In the preceding section we have shown that in the case of a pair of vectors $(\vec{\imath}, \vec{p})$ the probability $p_o \equiv p_{o,2}$ is approximately a power law; see Eq. (57). When dealing with N_{var} random variables represented by vectors $\vec{\imath}, \vec{p}, \vec{q}, \vec{r}, \ldots$ there are N_c different entries in the correlation matrix; see Eq. (2). The probability that each such entry is zero equals p_o. In order to roughly estimate results for N_{var} random variables, we now assume that the correlation matrix entries (random variables) are independent. They are only pairwise independent but we can assume that they are independent (see the discussion in Section 4.4 for the justification for this assumption). Thus, the joint probability mass function of all correlation coefficients (corresponding to all N_c pairs) is the product of marginal probability mass functions. Therefore, the probability that all N_c entries equal zero can be approximated as

$$p_{o,N_{\mathrm{var}}} \gtrapprox (p_o)^{N_c} \approx \left(\frac{12}{\sqrt{2\pi}} N_{\mathrm{sim}}^{-5/2} \right)^{\binom{N_{\mathrm{var}}}{2}}. \tag{62}$$

The probability of finding an optimal solution is also simply the ratio between the number of optimal solutions and the number of all possible solutions:

$$p_{o,N_{\mathrm{var}}} = \frac{n_{o,N_{\mathrm{var}}}}{(N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1}}. \tag{63}$$

A combination of Eqs. (62) and (63) gives us an estimate of the number of optimal solutions for arbitrary sample size N_{sim} and dimension N_{var} as a product

$$\begin{aligned} n_{o,N_{\mathrm{var}}} &\approx (N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1} \times p_{o,N_{\mathrm{var}}} \\ &\approx (N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1} \times \left(\frac{12}{\sqrt{2\pi}} N_{\mathrm{sim}}^{-5/2} \right)^{\binom{N_{\mathrm{var}}}{2}}. \end{aligned} \tag{64}$$

For a fixed N_{var}, the first factor grows faster with N_{sim} than the second factor decreases and therefore $n_{o,N_{\mathrm{var}}}$ grows with N_{sim}. The growth is faster for greater N_{var}. The above approximations for $n_{o,N_{\mathrm{var}}}$ and $p_{o,N_{\mathrm{var}}}$ are plotted in Fig. 7 and compared with the exact data from Table 2 (corrected by Eq. (61)). The figure shows that the probabilities $p_{o,N_{\mathrm{var}}}$ can be reasonably approximated by power laws (Eq. (62)), i.e. $p_{o,N_{\mathrm{var}}} \propto N_{\mathrm{sim}}^{-\frac{5}{4}(N_{\mathrm{var}}{}^2 - N_{\mathrm{var}})}$; see the dashed lines.

The agreement between the exactly computed values and approximations is reasonably good; Eq. (64) provides a good lower bound for the probabilities and for the numbers of optimal solutions. The main result is the implication that with increasing sample size N_{sim} the number of optimal solutions explodes; the growth is even faster for greater problem dimension N_{var}. Of course, a combination of very small N_{sim} with large N_{var} may result in the nonexistence of any optimal solution; see the development of the dashed lines in Fig. 7 (left).

The conditions for the existence of (at least one arrangement of) mutually Spearman-uncorrelated vectors can obtained by postulating that the number of optimal solutions is greater than one: $n_{o,N_{\mathrm{var}}} > 1$ in Eq. (64). For larger values of both N_{sim} and N_{var} this can be simplified using Stirling's approximation into

$$\frac{12}{\sqrt{2\pi}} \exp\left(-2 \frac{N_{\mathrm{sim}}}{N_{\mathrm{var}}} \right) \times N_{\mathrm{sim}}^{\frac{2N_{\mathrm{sim}}+1}{N_{\mathrm{var}}} - \frac{5}{2}} > 1. \tag{65}$$

This yields to a condition that $N_{\mathrm{sim}} \gtrapprox \frac{3}{2} N_{\mathrm{var}}$ for small N_{var}. Analysis of the leading terms yields the limiting condition for large N_{var}:

$$N_{\mathrm{sim}} > \frac{5}{4} N_{\mathrm{var}}. \tag{66}$$

Unfortunately, as will be shown in part III, the algorithm proposed in part I is unable to find any of these optimal solutions in a reasonable time whenever $N_{\mathrm{var}} > 4$, although the same holds for other algorithms studied there. However, the above-described procedure for finding vectors of ranks leading to zero intercorrelations suggests that the N_{var} dimensional combinatorial optimization problem may be more efficiently solved by a sequence of $N_{\mathrm{var}} - 1$ significantly smaller problems. In particular, one can progressively optimize problems with maximum $N_{\mathrm{sim}}!$ different correlations between a pair by permuting variables $\vec{p}$ against $\vec{\imath}$, then a triple of variables by permuting $\vec{q}$ against $\{\vec{\imath}, \vec{p}\}$, then a quaternion of variables by permuting $\vec{r}$ against $\{\vec{\imath}, \vec{p}, \vec{q}\}$, etc. Solving $N_{\mathrm{var}} - 1$ problems which have $N_{\mathrm{sim}}!$ possible results each seems to be more efficient than solving one problem with $(N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1}$ possible results.

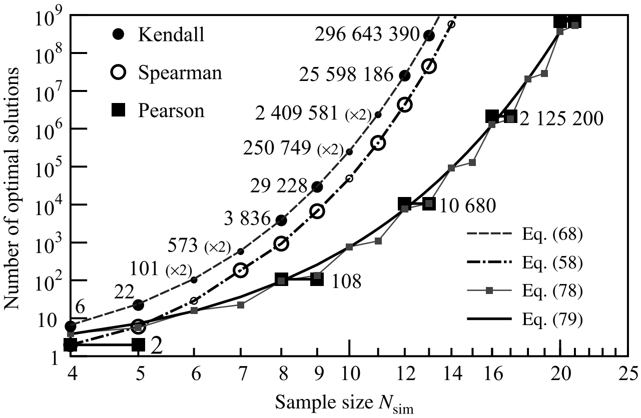

Fig. 8. Optimal solutions for two random variables.

5.2. Kendall-uncorrelatedness

5.2.1. Pair of random variables

In analogy with Section 5.1.1, we will now study the number of optimal solutions leading to a Kendall correlation of zero or the one given in Eq. (21). The number n_o grows even faster with N_{sim} than in the case of the Spearman correlation coefficient — compare the numbers in Table 2 with the numbers corresponding to Kendall correlation in Fig. 8.

Again, the estimate of n_o can be deduced from the fact that the $p_o/\Delta t$ converge to the peak of Gaussian pdf that approximates the exact sampling distribution $f_E(t)$:

$$\begin{aligned} f_o &= \frac{p_o}{\Delta t_{\mathrm{Kendall}}} = \frac{n_{o,\mathrm{Kendall}}}{N_{\mathrm{sim}}!} \frac{1}{\Delta t_{\mathrm{Kendall}}} \\ &\approx f_G(0) = \frac{1}{\sigma_t^{\mathrm{Kendall}} \sqrt{2\pi}} = \frac{3}{2} \sqrt{\frac{N_{\mathrm{sim}}\,(N_{\mathrm{sim}} - 1)}{2\pi\ \left(N_{\mathrm{sim}} + \frac{5}{2}\right)}}. \end{aligned} \tag{67}$$

Therefore, the number of optimal solutions is approximately

$$n_{o,\mathrm{Kend}} \approx \frac{6\,N_{\mathrm{sim}}!}{\sqrt{2\pi N_{\mathrm{sim}}\,(N_{\mathrm{sim}} - 1)\left(N_{\mathrm{sim}} + \frac{5}{2}\right)}} \approx \frac{6\,N_{\mathrm{sim}}!}{\sqrt{2\pi}\,N_{\mathrm{sim}}^{3/2}}. \tag{68}$$

The quality of the approximation can be assessed using Fig. 8, where the exact numbers of solutions (written close to the solid circles) are compared to the above equation (dashed line). The figure presents exact numbers for N_{sim} up to 13. When $N_{\mathrm{sim}} = 14$ the exact result is (2×)3 727 542 188 and for $N_{\mathrm{sim}} = 15$ an exhaustive search gives (2×)50 626 553 988. The error of approximation by Eq. (68) decreases with increasing N_{sim}. For $N_{\mathrm{sim}} = 4$ the error is 8.4% and for $N_{\mathrm{sim}} = 15$ the error drops below 2% (the approximation overestimates the exact results). Another way of approximating the number of optimal solutions is:

$$n_{o,\mathrm{Kend}} \approx \frac{6\,\Gamma\,(N_{\mathrm{sim}})}{\sqrt{2\pi N_{\mathrm{sim}}}} \approx 6 \frac{N_{\mathrm{sim}}^{N_{\mathrm{sim}}-1}}{\exp\,(N_{\mathrm{sim}})}. \tag{69}$$

Comparison of Eqs. (58) and (68) reveals how many times the number of Kendall optimal solutions is greater than the number of Spearman optimal solutions; see Fig. 8:

$$\frac{n_{o,\mathrm{Kend}}}{n_{o,\mathrm{Spear}}} \approx \frac{N_{\mathrm{sim}}}{2}. \tag{70}$$

5.2.2. Three and more random variables

The extension of the estimate of a number of optimal solutions into more dimensions can be performed the same way as in Section 5.1.2. We first estimate the probability of hitting an optimal solution (unit correlation matrix) when permuting N_{var} vectors as

$$p_{o,N_{\mathrm{var}}} = \frac{n_{o,N_{\mathrm{var}}}}{(N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1}} \lessapprox (p_o)^{N_c} \approx \left(\frac{6}{\sqrt{2\pi}} N_{\mathrm{sim}}^{-3/2}\right)^{\binom{N_{\mathrm{var}}}{2}}. \tag{71}$$

This gives us an estimate of the number of optimal solutions for arbitrary sample size N_{sim} and dimension N_{var} as a product (compare with Eq. (64)):

$$n_{o,N_{\mathrm{var}}} \approx (N_{\mathrm{sim}}!)^{N_{\mathrm{var}}-1} \times \left(\frac{6}{\sqrt{2\pi}} N_{\mathrm{sim}}^{-3/2}\right)^{\binom{N_{\mathrm{var}}}{2}}. \tag{72}$$

Again, for a fixed N_{var}, the first factor grows faster with N_{sim} than the second factor decreases and therefore $n_{o,N_{\mathrm{var}}}$ grows with N_{sim}. The growth is also faster for greater N_{var}.

As in the case of Spearman correlation, the question arises as to whether a combination of very small N_{sim} with large N_{var} results in the nonexistence of any optimal solution. The conditions for the existence of (at least one arrangement of) mutually Kendall-uncorrelated vectors can obtained by postulating that the number of optimal solutions is greater than one: $n_{o,N_{\mathrm{var}}} > 1$ in Eq. (72). For larger values of both N_{sim} and N_{var} this can be simplified using Stirling's approximation into

$$\frac{6}{\sqrt{2\pi}} \exp\left(-2\frac{N_{\mathrm{sim}}}{N_{\mathrm{var}}}\right) \times N_{\mathrm{sim}}^{\left(\frac{2N_{\mathrm{sim}}+1}{N_{\mathrm{var}}} - \frac{3}{2}\right)} > 1. \tag{73}$$

This yields to a condition that $N_{\mathrm{sim}} \gtrapprox N_{\mathrm{var}}$ for small N_{var}. The exponent suggests the limiting condition for large N_{var}: $N_{\mathrm{sim}} > \frac{3}{4}N_{\mathrm{var}}$. It is easy to show that to estimate a positive definite correlation matrix (particularly the unit matrix), the sample size N_{sim} must exceed the dimension N_{var}. Therefore, we conclude that the optimality can only be achieved when $N_{\mathrm{sim}} > N_{\mathrm{var}}$.

5.3. Optimal solutions yielding Pearson-uncorrelatedness of LHS-sampled Gaussian variables

5.3.1. Pair of random variables

We now construct all possible vectors orthogonal to a given vector of the dimension $N_{\mathrm{sim}} = 0 \pmod 4$. Vectors of a dimension equal to the sample size increased by one can be constructed by adding the zero center points. Once again, the total number of orthogonal vectors will be termed n_o. We construct orthogonal vectors with arbitrary coordinates (real values or integers) and therefore such constructions can also be used for Spearman-uncorrelated vectors.

The dot product in Eq. (6) needed for the computation of correlation sums N_{sim} products of pairs of values. What we need in order to surely achieve zero correlation (orthogonality of vectors) is to construct $m = N_{\mathrm{sim}}/2$ products with positive signs (composed of two positive or two negative values) which are each balanced by one of the $N_{\mathrm{sim}}/2$ identical products with negative signs. The m positive values in each vector can be represented by symbols $\{a, b, c, d, e \ldots\}$. We will collect them into a set, $\boldsymbol{u}$. Since $N_{\mathrm{sim}} = 0 \pmod 4)$, m is even and thus divisible by two.

Let us now construct all possible vectors $\boldsymbol{v}$ by permuting the values (coordinates) from $\boldsymbol{u}$. The dot product $\boldsymbol{u}\cdot\boldsymbol{v}$ sums m products. For example, when $N_{\mathrm{sim}} = 8$ the half length $m = 4$ and the first vector might look like $\boldsymbol{u} = \{a, b, c, d\}$ while the permuted vector may look like $\boldsymbol{v} = \{b, c, d, a\}$. The original vector and the vector orthogonal to it are constructed from the two half-length vectors by conveniently assigning the two signs + and − so that the four products $\{ab, bc, cd, da\}$ will be canceled by similar pairs composed from the other halves of the samples. Note that the above permutation $\boldsymbol{v}$ of vector $\boldsymbol{u}$ is a *cyclic permutation* with a cycle length $i = m$ (symbol a from the first vector maps b in the second one from which the first vector points to c, etc., and the loop is closed after four references). Note that the permutation $\boldsymbol{v}$ of $\boldsymbol{u}$ cannot have a fixed point (no symbol can be mapped to itself) because their product cannot be canceled in any way. Moreover, one can easily show that the permutation of the symbols must consist solely of *cycles with even lengths i* in order to be used in the construction of orthogonal vectors. A special case of such a cycle with an even length $i = 2$ is called a transposition. The longest possible cycle can be of length $i = m$ and there are $(m-1)!$ such permutations possible. Let us use term p_i to denote the number of cycles of length i. In order to construct a permutation from even cycles i with the multiplicity p_i, the vector length m must be partitioned into even numbers i so that

$$\sum_{i=2}^{m} i\, p_i = m = \frac{N_{\mathrm{sim}}}{2}. \tag{74}$$

The number of different ways to partition m into even numbers i is equal to the number of partitions of $m/2$ into positive integers (a well known and solved problem). Let us call a given partition a C *composition*. The number of ways to construct a given type C composition is equal to

$$M(C) = \frac{m!}{\prod_{2}^{m} \left(i^{p_i}\, p_i!\right)}. \tag{75}$$

Each composition must be fitted with the two signs in such a way that the products in $\boldsymbol{u}\cdot\boldsymbol{v}$ will be canceled. The number of convenient ways to distribute the signs to p_i cycles of length i will be called s_i from here on. When the cycle length $i = 2$, there are only two ways to do it and therefore $s_2 = 2^{p_i}$. This is anomalous and it corresponds to the two pairs of type ±$a \times b$ when $N_{\mathrm{sim}} = 4$ (see Table 1 in Section 3.3). In the rest of the cases there are $s_i = 2^{i\,p_i}$ ways to distribute the signs between all p_i cycles of an even length i. Note that when there is no cycle of such length $(p_i) = 0$, the number $s_i = 1$. This is an important feature because the total number of ways to distribute signs in a given type C composition is

$$S(C) = \prod_{2}^{m} s_i. \tag{76}$$

The total number of different constructions of orthogonal vectors of type C composition equals

$$N(C) = M(C) \times S(C). \tag{77}$$

Finally, in order to count all possible constructions of orthogonal vectors n_o one has to sum $N(C)$ for all different C compositions. For example, when $N_{\mathrm{sim}} = 12 \Leftrightarrow m = 6$, there are three types of compositions, namely one large cycle of length $\{6\}$, two cycles of lengths $\{4, 2\}$ and three cycles of lengths $\{2, 2, 2\}$ (these three compositions correspond to a partitioning of the number three into 3 or 2, 1 or 1, 1, 1). Table 4 shows the computation.

It can be noticed that majority of the orthogonal vectors are constructed using compositions with the longest possible cycle of length $i = m$. The number of such solutions reads

$$N(m) = (m-1)! \times 2^m. \tag{78}$$

Even though the proportion $N(m)/n_o$ decreases with increasing m, one can use $N(m)$ as a sure lower bound on the number n_o giving roughly its order of magnitude (the fraction is about 89% when $N_{\mathrm{sim}} = 8$; 74% when $N_{\mathrm{sim}} = 12$ (see Table 4); 61% when $N_{\mathrm{sim}} = 16$ and 54% when $N_{\mathrm{sim}} = 20$). Therefore, we can write

$$n_{o,\mathrm{Pears}} \gtrapprox N(m) \approx \frac{N_{\mathrm{sim}}^{\frac{N_{\mathrm{sim}}-1}{2}}}{\exp\left(\frac{1}{2}N_{\mathrm{sim}}\right)} 2\sqrt{\pi}. \tag{79}$$

Fig. 8 compares the numbers of pairs of Pearson-uncorrelated vectors (orthogonal vectors) with the numbers of Spearman-

Table 4
Computation of the number of $n_{o,\mathrm{Pears}} = 10{,}680$ orthogonal vectors when $N_{\mathrm{sim}} = 12 \Leftrightarrow m = 6$.

Compos. type	$\{6\}$		$\{4, 2\}$		$\{2, 2, 2\}$	
Cycle length i	p_i	s_i	p_i	s_i	p_i	s_i
2	0	1	1	2	3	2^3
4	0	1	1	2^4	0	1
6	1	2^6	0	1	0	1
$M(C), S(C)$	5!	2^6	90	2^5	15	8
$N(C)$	7860		2880		120	
$n_{o,\mathrm{Pears}}$ (sum)	10,680					

uncorrelated vectors counted in Section 5.1. It is seen that Kendall correlation allows for the biggest number, n_o. It is also clear that in the Spearman case, the number n_o is greater compared to the number of orthogonal vectors $n_{o,\mathrm{Pears}}$ at the same sample size N_{sim}. Why? Orthogonality with real numbers is a stricter requirement than Spearman or Kendall-uncorrelatedness obtained with integer ranks. To highlight the difference between the numbers of Spearman and Pearson correlations, we now square $n_{o,\mathrm{Pears}}$ from Eq. (79) and divide it by N_{sim}:

$$\frac{n_{o,\mathrm{Pears}}^2}{N_{\mathrm{sim}}} \approx 4\pi \frac{N_{\mathrm{sim}}^{N_{\mathrm{sim}}-2}}{\exp\left(N_{\mathrm{sim}}\right)} \approx n_{o,\mathrm{Spear}}. \tag{80}$$

Comparison with Eq. (59) immediately shows that the number of Spearman orthogonal solutions is approximately equal to the square of Pearson solutions divided by N_{sim}.

5.3.2. Three and more random variables

The situation is more complicated when constructing three or more mutually orthogonal vectors. For example, when $N_{\mathrm{sim}} = 12$ (or 13), no third vector that is orthogonal to the previous two can be constructed to any of the 10 680 pairs from Table 4. Surprisingly, for a smaller sample size many solutions exist. In particular, when $N_{\mathrm{sim}} = 8$ (or 9), there are $n_{o,3} = 216$ orthogonal triples. The exhaustive search based on the construction of a pool of orthogonal vectors to a given vector using the method suggested in the previous section and then selecting the third, fourth, etc. vector from the pool quickly becomes too expensive, which renders it useless.

In Sections 5.1.1 and 5.2.1, which are focused on Spearman and Kendall correlations, we have exploited the knowledge of the distribution of a random correlation coefficient to estimate $n_{o,N_{\mathrm{var}}}$ for arbitrary N_{var}. The key fact used there is that Spearman and Kendall correlations take on values (troughs) that are uniformly distributed over the interval $\langle -1, 1\rangle$. The probability density could therefore be readily transformed into a number of occurrences (multiplicity) of particular correlations. Unfortunately, the attainable Pearson correlations are not distributed uniformly over the correlation range and we cannot apply the same trick. We conclude by referring to the papers on orthogonal arrays mentioned at the beginning of Section 5 where the constructions are solved for selected N_{var} and N_{sim}. These papers prove that mutually orthogonal vectors *occur* in higher dimensions, at least for those selected combinations of N_{sim} and N_{var}.

6. Conclusions

The paper presents theoretical (analytical) and some numerical results regarding the bounds of correlation errors for estimated correlation matrices. The results are focused on Spearman and Kendall rank order correlations, as well as Pearson linear correlation. The Spearman and Kendall correlation coefficients are distribution free and therefore the results are of wide applicability. All the results presented for Pearson's correlation are stated for Gaussian variables sampled via the LHS method (in two alternatives described in Part I, namely LHS-mean and LHS-median). Some of the results for Pearson's correlation can be applied to any symmetrical distribution sampled via the two methods. The most important results are as follows.

- Analysis of the minimum error in sampling correlation for Spearman, Pearson and Kendall correlation coefficients when an arbitrary correlation coefficient is requested. The analysis is based on the structure of the formula for correlation estimation. Special attention has been paid to the lower bound on correlation error when uncorrelatedness is targeted. These exact errors are derived for all three correlation coefficients.
- Derivation of the possibility and number of optimal solutions, i.e. mutual orderings of vectors representing random variables whose pairwise correlations are zero or attain the lower bounds from the preceding item.
- Statistical distribution of the correlation norms ρ_{rms} and $\rho_{\max}$ defined in Part I as they occur when sample ordering is left random. These norms serve as the upper bound on the efficiency of the combinatorial algorithm proposed in Part I. It is a common practice to model uncorrelated or even independent random variables using vectors whose ordering is shuffled independently. In such cases the knowledge of the distributions is also important.

These results are used in the companion paper Part III which presents the performance of the correlation induction algorithm presented in Part I.

Special attention must be paid to situations when the sample size N_{sim} is smaller than the number of variables N_{var}. In these cases, the estimated correlation matrix is singular, yet the analyst may request the best possible match between the target and actual correlations. The cases when $N_{\mathrm{sim}} \leq N_{\mathrm{var}}$ are thoroughly analyzed in [52].

Acknowledgments

This outcome has been achieved with the financial support of the Grant Agency of the Academy of Sciences of the Czech Republic under project No. KJB201720902 and from the Czech Science Foundation under project No. P105/11/1385. The support is gratefully acknowledged.

The author thanks his friend, the mathematician Tomáš Hanžl, for his numerous valuable discussions and suggestions.

References

[1] Vořechovský M, Novák D. Correlation control in small sample Monte Carlo type simulations I: a simulated annealing approach. Probabilistic Engineering Mechanics 2009;24(3):452–62.
[2] Owen AB. Controlling correlations in Latin hypercube samples. Journal of the American Statistical Association Theory and Methods 1994;89(428):1517–22.
[3] Wang Y-T, Lam F, Barrett JD. Simulation of correlated modulus of elasticity and compressive strength of lumber with gain factor. Probabilistic Engineering Mechanics 1995;10(2):63–71.
[4] Huntington DE, Lyrintzis CS. Improvements to and limitations of Latin hypercube sampling. Probabilistic Engineering Mechanics 1998;13(4):245–53.
[5] Vořechovský M, Novák D. Statistical correlation in stratified sampling. In: Der Kiureghian A, Madanat S, Pestana JM, editors. ICASP 9, International Conference on Applications of Statistics and Probability in Civil Engineering. Rotterdam (Netherlands): Millpress; 2003. p. 119–24.
[6] Kendall MG. A new measure of rank correlation. Biometrika 1938;30(1–2): 81–9.
[7] Knight WR. A computer method for calculating Kendall's tau with ungrouped data. Journal of the American Statistical Association 1966;61(314):436–9.
[8] Daniels HE. Rank correlation and population models. Journal of the Royal Statistical Society. Series B Methodological 1950;12(2):171–91.
[9] Durbin J, Stuart A. Inversions and rank correlation coefficients. Journal of the Royal Statistical Society. Series B Methodological 1951;13(2):303–9.

[10] Daniels HE. Note on Surbin and Stuart's formula for E(rs). Journal of the Royal Statistical Society. Series B Methodological 1951;13(2):310.

[11] Genest C, Nešlehová J. Analytical proofs of classical inequalities between Spearman's ρ and Kendall's τ. In: The 8th Tartu conference on multivariate statistics & the 6th conference on multivariate distributions with fixed marginals. Journal of Statistical Planning and Inference 2009;139(11):3795–8. [special issue].

[12] Nelsen RB. On measures of association as measures of positive dependence. Statistics & Probability Letters 1992;14(4):269–74.

[13] Kendall MG. The advanced theory of statistics, vol. 1. 4th ed. Charles Griffin & Co. Ltd; 1948.

[14] Gibbons JD. Nonparametric methods for quantitative analysis. Holt McDougal; 1976.

[15] Fredricks GA, Nelsen RB. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. Journal of Statistical Planning and Inference 2007;137(7):2143–50.

[16] Li X, Li Z. Proof of a conjecture on Spearman's [rho] and Kendall's [tau] for sample minimum and maximum. Journal of Statistical Planning and Inference 2007;137(1):359–61.

[17] Chen Y-P. A note on the relationship between Spearman's [rho] and Kendall's [tau] for extreme order statistics. Journal of Statistical Planning and Inference 2007;137(7):2165–71.

[18] Vořechovský M. Correlation control in small sample Monte Carlo type simulations III: Performance study, multivariate modeling and copulas. Probabilistic Engineering Mechanics 2011; [in review].

[19] Iman RC, Conover WJ. Small sample sensitivity analysis techniques for computer models with an application to risk assessment. Communications in Statistics: Theory and Methods 1980;A9(17):1749–842.

[20] Pearson K. On further methods of determining correlation. Journal of the Royal Statistical Society 1907;70(4):655–6.

[21] Hotelling H, Pabst MR. Rank correlation and tests of significance involving no assumption of normality. The Annals of Mathematical Statistics 1936;7(1): 29–43.

[22] Pitman EJG. Significance tests which may be applied to samples from any populations. II, the correlation coefficient test. Journal of the Royal Statistical Society 1937;4(2):225–32. [Supplement].

[23] Olds EG. Distributions of sums of squares of rank differences for small numbers of individuals. The Annals of Mathematical Statistics 1938;9(2):133–48.

[24] David ST, Kendall MG, Stuart A. Some questions of distribution in the theory of rank correlation. Biometrika 1951;38(1–2):131–40.

[25] de Jonge C, van Montfort A. The null distribution of Spearman's s when $n = 12$. Statistica Neerlandica 1972;26(1):15–7.

[26] Zar JH. Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association 1972;67(339):578–80.

[27] Otten A. The null distribution of Spearman's $\underline{s}$ when n = 13(1)16. Statistica Neerlandica 1973;27(1):19–20.

[28] Otten A. Note on the Spearman rank correlation coefficient. Journal of the American Statistical Association 1973;68(343):585.

[29] Best D, Roberts D. Algorithm AS 89: the upper tail probabilities of Spearman's rho. Applied Statistics 1975;24(3):377–9.

[30] Franklin LA. A note on approximations and convergence in distribution for Spearman's rank correlation coefficient. Communications in Statistics – Theory Methods 1988;17(1):55–9.

[31] Franklin LA. The complete exact null distribution of Spearman's rho for $n = 12(1)18$. Journal of Statistical Computation and Simulation 1988;29(3): 578–80.

[32] Franklin LA. A note on the Edgeworth approximation to the distribution of Spearman's Rho with a correction to Pearson's approximation. Communications in Statistics – Simulation and Computation 1989;18(1):245–52.

[33] Ramsey PH. Critical values for Spearman's rank order correlation. Journal of Educational Statistics 1989;14(3):245–53.

[34] van de Wiel MA, Bucchianico AD. Fast computation of the exact null distribution of Spearman's ρ and page's L statistic for samples with and without ties. Journal of Statistical Planning and Inference 2001;92(1–2): 133–45.

[35] Kowalski CJ. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. Journal of the Royal Statistical Society. Series C Applied Statistics 1972;21(1):1–12.

[36] Fisher RA. Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. Biometrika 1915;10(4): 507–21.

[37] Rousseeuw PJ, Molenberghs G. The shape of correlation matrices. The American Statistician 1994;48(4):276–9.

[38] Gnedenko BV. Sur la distribution limite du terme maximum d'une série aléatorie. Annals of Mathematics 2nd Ser 1943;44(3):423–53.

[39] Gumbel EJ. Statistics of Extremes. New York: Columbia University Press; 1958.

[40] Fisher RA, Tippett LHC. Limiting forms of the frequency distribution of the largest and smallest member of a sample. Proceedings of the Cambridge Philosophical Society 1928;24:180–90.

[41] Fréchet M. Sur la loi de probabilité de l'écart maximum. Annales de la Sociéé Polonaise de Mathématique, Cracow 1927;6:93–122. printed in 1928.

[42] Castillo E. Extreme value theory in engineering. In: Statistical Modeling and Decision Science. London: Academic Press; 1988.

[43] Leadbetter MR, Lindgren G, Rootzen H. Extremes and related properties of random sequences and processes. In: Springer Series in Statistics. N.Y: Springer-Verlag; 1983.

[44] Owen AB. Orthogonal arrays for computer experiments, integration and visualization. Statistica Sinica 1992;2(2):439–52.

[45] Tang B. Orthogonal array-based Latin hypercubes. Journal of the American Statistical Association 1993;88(424):1392–7.

[46] McKay MD, Conover WJ, Beckman RJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.

[47] Ye KQ. Orthogonal column Latin hypercubes and their application in computer experiments. Journal of the American Statistical Association 1998;93(444): 1430–9.

[48] Butler N. Optimal and orthogonal Latin hypercube designs for computer experiments. Biometrika 2001;88(3):847–57.

[49] Steinberg DM, Lin DKJ. A construction method for orthogonal Latin hypercube designs. Biometrika 2006;10(2):279–88.

[50] Beattie S, Lin D. Rotated factorial design for computer experiments. In: Proceedings of Physical and Engineering Science section, 1997. American Statistical Association.

[51] Bursztyn D, Steinberg DM. Rotation designs: orthogonal first-order designs with higher order projectivity, In: Hirotsu Chihiro, Shinozaki Nobuo, editors. 2nd international symposium on business and industrial statistics, Yokohama, Japan. Applied Stochastic Models in Business and Industry 2001;18(3):20–1. [special issue].

[52] Vořechovský M. Optimal singular correlation matrices estimated when sample size is less than the number of random variables. Probabilistic Engineering Mechanics 2011. [in review].