Simulating the binary variates for the components of a socio-economical system

Ştefan V. Ştefănescu[1*]

[1]Research Institute for Quality of Life, Romanian Academy; stefanst@fmi.unibuc.ro

Abstract

Often in practice the components Wj of a sociological or an economical system $\underline{W}$ take discrete 0-1 values. We talk about how to generate arbitrary observations from a binary 0-1 system $\underline{B}$ when is known the multidimensional distribution of the discrete random vector $\underline{B}$. We also simulated a simplified structure of B given by the marginal distributions together with the matrix of the correlation coefficients. Different properties of the systems $\underline{W}$ are presented too.

Keywords: binary system, marginal distribution, Monte Carlo simulation, random variates, correlation coefficient.

1. Introduction

A general system $\underline{W}$ with k components $W_1, W_2, W_3, ..., W_k$ is characterized by the features λ_j of every variable W_j and the intensity c_{ij} of the relation between any two components W_i and W_j, $1 \le i, j \le k$. Frequently in practice the relation among the elements of the subsystem $\{W_i, W_j\}$ is a symmetric one, that is $c_{ij} = c_{ji}$.

The characteristic λ_j of the component W_j could be just the parameters which define the marginal distribution of the random variable W_j. In the following we will choose the Pearson correlation coefficient $Cor(W_i, W_j)$ to measure the intensity c_{ij} of the relation which is present between the components W_i and W_j of the system $\underline{W}$. We mention here that in the literature there are known many other indicators to measure the ratio among the elements W_i and W_j from $\underline{W}$ ([1], [2], [6]).

Figure 1 presents some kinds of systems $\underline{W}$.

Many times in practice the system $\underline{W}$ has components W_j with a normal distribution. Such a system will be designated in the subsequent by $\underline{X}$. For this particular case the system components X_j, $1 \le j \le k$, are dependent normal random variables characterized by their means μ_j and their dispersions σ_j^2. So we will take $\lambda_j = (\mu_j, \sigma_j)$ and $c_{ij} = Cor(X_i, X_j)$, $1 \le i, j \le k$.

Another class from the systems $\underline{W}$ are binary 0-1 systems designated by $\underline{B}$. The elements $B_1, B_2, B_3, ..., B_k$ of the system $\underline{B}$ are binary dependent variables which take only the values 0 and 1. To make a distinction between the systems $\underline{B}$ and $\underline{X}$ we will use the notation $r_{ij} = Cor(B_i, B_j)$ in the discrete case and $c_{ij} = Cor(X_i, X_j)$ for the continuous normal marginals variant.

We mention here that the normal type system $\underline{X}$ is completely characterized by the set of the parameters μ_i, σ_i, c_{ij}, $1 \le i < j \le k$, that is $k(k+3)/2$ values ([3]).

[*] Corresponding author: stefanst@fmi.unibuc.ro

But the multidimensional distribution of an arbitrary binary system $\underline{B}$ has more parameters. For this reason, in opposition with the normal distributions case, we can not define a general binary 0-1 system $\underline{B}$ by knowing only the values μ_i, σ_i, r_{ij}, $1 \le i < j \le k$. More, in the discrete case of $\underline{B}$, the variance $\sigma_j^2 = Var(B_j)$ depends on the mean $\mu_j = Mean(B_j)$. So, knowing only the marginals and the correlation matrix of $\underline{B}$ we lose a lot of information which define the real multivariate discrete distribution of the system $\underline{B}$. Some details concerning the behavior of a binary system $\underline{B}$ will be given in the next section.

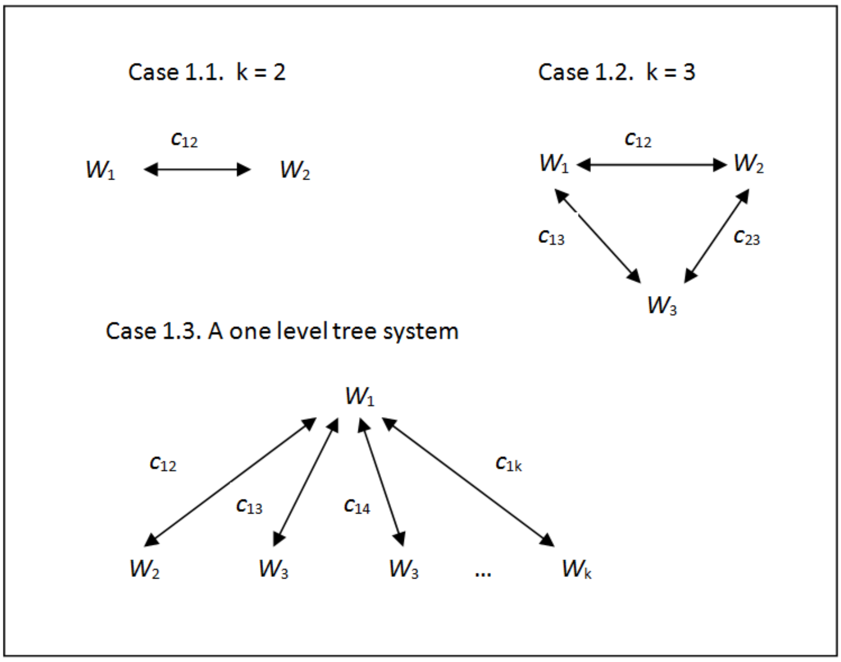


Fig. 1. A system $\underline{W}$ with k components

We reveal a new other aspect which is present for sociological and economical systems too. So, the individuals of a given population estimate the behaviour of each component W_j from a continuous system $\underline{W}$ by putting subjective marks.

In this approach a binary system $\underline{B}$ results from $\underline{W}$ when the marks take only 0 and 1 values. Hence, in practice, we often approximate a continuous system $\underline{W}$ by a binary one, like $\underline{B}$. In this case we must evaluate the discretization error.

2. The binary 0-1 systems

The binary random vector $\underline{B} = (B_1, B_2, B_3, \ldots, B_k)$ which takes only 0 and 1 values is completely characterized by the probabilities $p_{i_1,i_2,i_3,\ldots,i_k}$, $i_j \in \{0,1\}$, $1 \le j \le k$, where

$$p_{i_1,i_2,i_3,\ldots,i_k} = Pr(B_1 = i_1, B_2 = i_2, B_3 = i_3, \ldots, B_k = i_k)$$

Obviously, $p_{i_1,i_2,i_3,\ldots,i_k} \ge 0$ for all indices $i_j \in \{0,1\}$ and in addition

$$\sum_{i_1=0}^{i_1=1} \sum_{i_2=0}^{i_2=1} \sum_{i_3=0}^{i_3=1} \cdots \sum_{i_k=0}^{i_k=1} p_{i_1,i_2,i_3,\ldots,i_k} = 1 \qquad (1)$$

To simplify our expose, for any $i_j \in \{0,1\}$, we will use the notation

$p_{i_1,\ldots,i_{j-1},+,i_{j+1},\ldots,i_k} = p_{i_1,\ldots,i_{j-1},0,i_{j+1},\ldots,i_k} + p_{i_1,\ldots,i_{j-1},1,i_{j+1},\ldots,i_k}$

So, the equality (1) could be also written in a shorter form as $p_{+,+,+,\ldots,+} = 1$.

The marginal distributions of the random vector $\underline{B}$ are defined only by the probabilities $q_j = Pr(B_j = 1)$, $1 \le j \le k$.

Choosing, for example, the component B_1 we deduce

$Pr(B_1 = 0) = p_{0,+,+,\ldots,+} = 1 - p_{1,+,+,\ldots,+} = 1 - Pr(B_1 = 1) = 1 - q_1$

Remark 1. Since the distribution of the system $\underline{B} = (B_1, B_2, B_3, \ldots, B_k)$ is determined by the probabilities $p_{i_1,i_2,i_3,\ldots,i_k}$ with the restriction (1) we conclude that a general binary 0-1 system $\underline{B}$ with k components is defined by $2^k - 1$ parameters.

Now we will enumerate some properties of a binary $\underline{B} = (B_1, B_2)$ system which has only two components.

We remind that the distribution of an arbitrary 0-1 binary vector $\underline{B} = (B_1, B_2)$ is given by the probabilities $p_{i,j} = Pr(B_1 = i, B_2 = j)$ where $i, j \in \{0, 1\}$ and $p_{+,+} = 1$

In this case $q_1 = p_{1,+} = Pr(B_1 = 1)$, $q_2 = p_{+,1} = Pr(B_2 = 1)$, $0 \le q_1, q_2 \le 1$ and therefore

$p_{1,0} = q_1 - p_{1,1}$, $p_{0,1} = q_2 - p_{1,1}$, $p_{0,0} = 1 + p_{1,1} - q_1 - q_2$

Hence we have the inequalities

P2.1. $max\{0, q_1 + q_2 - 1\} \le min\{q_1, q_2\}$

After a straightforward calculus we obtain the relations

P2.2. $Mean(B_j) = Mean(B_j^2) = q_j$, $Var(B_j) = q_j(1 - q_j)$, $j \in \{0, 1\}$

$$r_{12} = Cor(B_1, B_2) = \frac{p_{1,1} - q_1 q_2}{\sqrt{q_1(1-q_1)}\,\sqrt{q_2(1-q_2)}}, \quad 0 < q_1, q_2 < 1$$

Remark 2. This expression of the correlation coefficient $r_{12} = Cor(B_1, B_2)$ does not depend on the concrete values of the binary random variables B_1 and B_2. For example, considering $B_1 \in \{a_1, b_1\} \ne \{0, 1\}$, $B_2 \in \{a_2, b_2\} \ne \{0, 1\}$ we obtain the same value for the indicator r_{12}.

Since $q_1 = p_{1,0} + p_{1,1}$ and $q_2 = p_{0,1} + p_{1,1}$ we prove easily

P2.3. If $p_{1,1} = q_1 q_2$ then we have also the following equalities

$p_{0,1} = (1 - q_1) q_2$, $p_{1,0} = q_1(1 - q_2)$, $p_{0,0} = (1 - q_1)(1 - q_2)$

From P2.2 and P2.3 it results

P2.4. The binary 0-1 random variables B_1, B_2 are independent if and only if $r_{12} = Cor(B_1, B_2) = 0$.

Remark 3. The property P2.4 is not always true for an arbitrary continuous two component system $\underline{W} = (W_1, W_2)$.

Applying the propositions P2.1 and P2.2 we deduce the inequalities

P2.5. $$Cor(B_1, B_2) \ge \frac{max\{0, q_1 + q_2 - 1\} - q_1 q_2}{\sqrt{q_1(1-q_1)}\,\sqrt{q_2(1-q_2)}}, \quad 0 < q_1, q_2 < 1$$

$$Cor(B_1,B_2) \le \frac{min\{q_1, q_2\} - q_1 q_2}{\sqrt{q_1(1-q_1)}\sqrt{q_2(1-q_2)}}, \quad 0 < q_1, q_2 < 1$$

The following properties are particular cases of the proposition P2.5.

P2.6. If $q_1 = q_2$ then $Cor(B_1,B_2) \le 1$

If $q_1 = 1 - q_2$ then $Cor(B_1,B_2) \ge -1$

Using the formulas

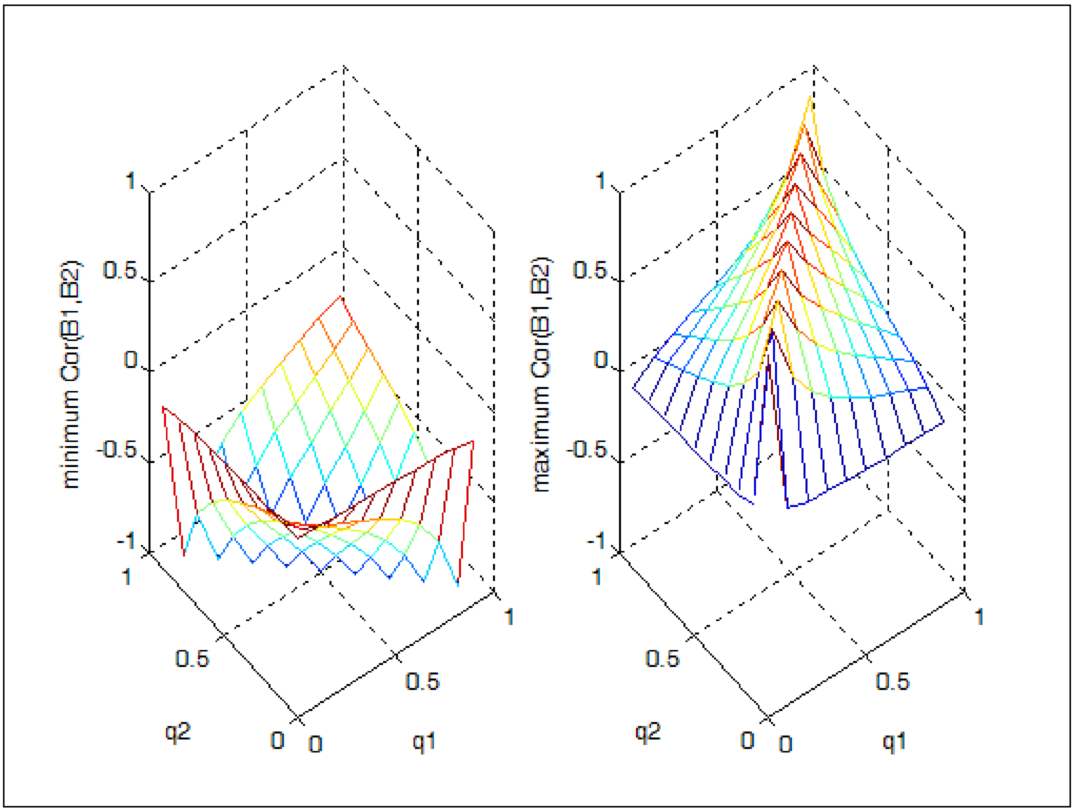
$$Cov(1-B_1, B_2) = -Cov(B_1, B_2), \quad Var(1-B_1, B_2) = Var(B_1, B_2)$$

we can prove directly the equalities

P2.7. $Cor(1-B_1, B_2) = Cor(B_1, 1-B_2) = -Cor(B_1, B_2)$

Graphic 1 presents us a suggestive image of the variation for the lower and upper bounds of $r_{12} = Cor(B_1,B_2)$ index depending on the marginal distributions indicators $0 < q_1, q_2 < 1$.

Remark 4. From the propositions P2.1-P2.7 we conclude that the discrete distribution of the system $\underline{B} = (B_1, B_2)$ is completely determined by the indices $0 < q_1, q_2 < 1$ which characterize the marginal distributions of $\underline{B}$ together with the correlation coefficient $r_{12} = Cor(B_1,B_2)$, $-1 \le r_{12} \le 1$. But the parameters q_1, q_2, r_{12} are mutually dependent (see the properties P2.1 and P2.5 or *Graphic 1*).



Graphic 1. The lower and upper bounds of $r_{12} = Cor(B_1,B_2)$

3. Generate random observations from a binary system

Leisch, Weingessel and Hornik suggested in [5] the application of the general inverse method for discrete random vectors ([3], [4]) to generate arbitrary observations $(b_1, b_2, b_3, \dots, b_k)$, $b_j \in \{0,1\}$, for the system $\underline{B} = (B_1, B_2, B_3, \dots, B_k)$.

The following algorithm *GDRV* produces $(b_1, b_2, b_3, \dots, b_k)$ vectors, $b_j \in \{0,1\}$, such that

$$Pr(B_1 = b_1, B_2 = b_2, B_3 = b_3, \dots, B_k = b_k) = p_{b_1, b_2, b_3, \dots, b_k}$$

where the probabilities $p_{i_1, i_2, i_3, \dots, i_k}$, $i_j \in \{0,1\}$, $1 \le j \le k$, define the binary 0-1 system $\underline{B}$.

Algorithm GDRV (Generating Discrete Random Vectors).

Step 0. Input : the probabilities $p_{i_1, i_2, i_3, \dots, i_k}$, $i_j \in \{0,1\}$, $1 \le j \le k$, with $p_{+,+,+,\dots,+} = 1$.

Step 1. Establish a one to function $h : \{1, 2, 3, \dots, 2^k\} \to \{0,1\}^k$

Step 2. Compute recurrently the sums

$$s_0 = 0$$

$$s_t = s_{t-1} + p_{h(t)}, \quad 1 \le t \le 2^k$$

Step 3. Generate a random variate u uniformly distributed on the interval $(0,1]$

Step 4. Find the index $1 \le t \le 2^k$ such that $u \in (s_{t-1}, s_t]$

Step 5. $b = h(t)$

Step 6. Output : b

Details regarding the theoretical justification of the generating procedure *GDRV* can be found in the books [3] and [4].

Remark 5. Applying algorithm *GDRV* we generated $n = 10^6$ random variates (b_1, b_2, b_3) from the binary system $\underline{B} = (B_1, B_2, B_3)$ defined by *Table 1*. For this case the frequences of the categories (i_1, i_2, i_3), $i_j \in \{0,1\}$, $1 \le j \le 3$, are given in *Table 2*. The validity of the algorithm *GDRV* is proved in part since the theoretical values and the empirical estimations of the probabilities p_{i_1, i_2, i_3} are very closed (compare the results from *Tables 1-2*).

Table 1. The theoretical distribution of the binary 0-1 system $\underline{B} = (B_1, B_2, B_3)$

$p_{0,0,0}$	$p_{0,0,1}$	$p_{0,1,0}$	$p_{0,1,1}$	$p_{1,0,0}$	$p_{1,0,1}$	$p_{1,1,0}$	$p_{1.1.1}$
0.050	0.200	0.100	0.150	0.100	0.050	0.050	0.300

Table 2. The frequences for the variates (b_1, b_2, b_3) ***obtained after*** 10^6 ***simulations with algorithm GDRV***

(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
49763	200067	99951	149842	99672	49832	50332	300541

4. Systems with normal distributed components

Now we will discuss the case of a system $\underline{X} = (X_1, X_2, X_3, ..., X_k)$ where its components X_j, $1 \le j \le k$, are random variables with normal distributions.

By $X \sim Norm(\mu, \sigma^2)$ with $\mu \in R$, $\sigma > 0$, we understand that the random variable X is normal distributed where $Mean(X) = \mu$ and $Var(X) = \sigma^2$. We denote by $\Phi(x)$ the Laplace function, that is the cumulative distribution function for the random variable $Z \sim Norm(0,1)$.

Remind some properties which will be applied in the subsequent.

P4.1. If $Z \sim Norm(0,1)$ and $X = \mu + \sigma Z$ with $\mu \in R$, $\sigma > 0$ then we have $X \sim Norm(\mu, \sigma^2)$.

P4.2 (Inverse method, [3], [4]). If the random variable U is uniformly distributed on the interval $[0,1]$ and $Z = \Phi^{-1}(U)$ then $Z \sim Norm(0,1)$.

P4.3. For any $\mu_i \in R$, $\sigma_i > 0$, if $X_i \sim Norm(\mu_i, \sigma_i^2)$ and $Y = X_1 + X_2$ then $Y \sim Norm(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Discretization procedure *DP*. For any $a \in R$, $\mu \in R$, $\sigma > 0$ and $X \sim Norm(\mu, \sigma^2)$ we designate by $B_{X,a}$ the following binary 0-1 random variable

$$B_{X,a} = \begin{cases} 0 & , \quad when \ X < a \\ 1 & , \quad when \ X \ge a \end{cases}$$

Using the procedure *DP* we deduce by a direct calculus

P4.4. For any $X \sim Norm(\mu, \sigma^2)$ we have $\Pr(B_{X.a} = 1) = 1 - \Phi((a - \mu)/\sigma)$

P4.5. For any $-1 \le c \le 1$, $Z_i \sim Norm(0,1)$, the standard normal random variables Z_1, Z_2 being independent, if

$$X = Z_1$$

$$Y = cZ_1 + \sqrt{1 - c^2}\, Z_2$$

then $X \sim Norm(0,1)$, $Y \sim Norm(0,1)$ and more $Cor(X, Y) = c$.

Remark 6. By using a normal random variable $X \sim Norm(\mu, \sigma^2)$ and a given bound $a \in R$ we build a binary 0-1 random variable $B_{X,a}$ such that

$$q = \Pr(B_{X.a} = 1) = 1 - \Phi((a - \mu)/\sigma)$$

(see the discretization procedure *DP* and *Proposition P4.4*). When $\mu = 0$ and $\sigma = 1$, the threshold $a \in R$ determine effectively the distribution of the discrete 0-1 random variable $B_{X,a}$.

5. A discretization process

Having a continuous normal distributed system $\underline{X} = (X_1, X_2, X_3, ..., X_k)$ and fixing some arbitrary thresholds $a_1, a_2, a_3, ..., a_k \in R$ we can obtain a binary 0-1 system $\underline{B} = (B_1, B_2, B_3, ..., B_k)$ with $B_j = B_{X_j, a_j}$, $1 \le j \le k$ (apply the procedure *DP*).

More, when $X_j \sim Norm(0,1)$, $1 \le j \le k$, then $q_j = \Pr(B_j = 1) = 1 - \Phi(a_j)$.

Obviously, in this last case, the correlation indicators $r_{ij} = Cor(B_i, B_j)$ and $c_{ij} = Cor(X_i, X_j)$, $1 \le i, j \le k$, have not equal values. More precisely, a correlation coefficient r_{ij} depends on the quantities c_{ij}, q_i, q_j. The effective relation between r_{ij} and c_{ij} indices will be established in the subsequent by applying a stochastic Monte Carlo simulation.

Remark 7. For an arbitrary $-1 \le c \le 1$, propositions *P4.2* and *P4.5* permit us to generate two dependent standard normal random variables X, Y having just the Pearson correlation coefficient $Cor(X,Y) = c$. We can apply *Proposition P4.2* (the inverse method, [3], [4]) to generate independent $Z_i \sim Norm(0,1)$ random variables which are used by *Proposition P4.5*.

Now, keeping all the previous notations, we will suggest a Monte Carlo procedure *MCRCC* to establish the real ratios between he correlation coefficients $c_{ij} = Cor(X_i, X_j)$ and $r_{ij} = Cor(B_i, B_j)$.

Procedure MCRCC.

Step 1. We generate random variates of volume n for a bidimensional random vector (X_1, X_2) with standard normal dependent marginals and $c_{12} = Cor(X_1, X_2)$, $-1 \le c_{12} \le 1$ (more details in *Remark 7*).

Step 2. Knowing the marginal probabilities $-1 \le q_1, q_2 \le 1$, we specify the discretization thresholds, that is $a_1 = \Phi^{-1}(1-q_1)$, $a_2 = \Phi^{-1}(1-q_2)$.

Step 3. We obtain 0-1 binary samples (b_1, b_2) from the random vector $\underline{B} = (B_i, B_j)$ considering the discretization procedure $B_1 = B_{X_1, a_1}$, $B_2 = B_{X_2, a_2}$ (algorithm *DP*).

Step 4. Using the samples resulted for $\underline{B} = (B_i, B_j)$ we estimate the correlation coefficient $r_{12} = Cor(B_1, B_2)$.

The correlation values r_{12} from *Tables 3-5* were deduced by running the Monte Carlo algorithm *MCRCC* for samples having the volume $n = 10^7$.

Table 3. $q_1 = q_2 = 0.5$, $n = 10^7$ ***Monte Carlo simulations with MCRCC***

c_{12}	-0.999	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4
r_{12}	-0.9714	-0.7129	-0.5906	-0.4938	-0.4099	-0.3335	-0.2621
c_{12}	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
r_{12}	-0.1940	-0.1282	-0.0637	0.0001	0.0638	0.1284	0.1943
c_{12}	0.4	0.5	0.6	0.7	0.8	0.9	0.999
r_{12}	0.2622	0.3333	0.4096	0.4937	0.5904	0.7129	0.9714

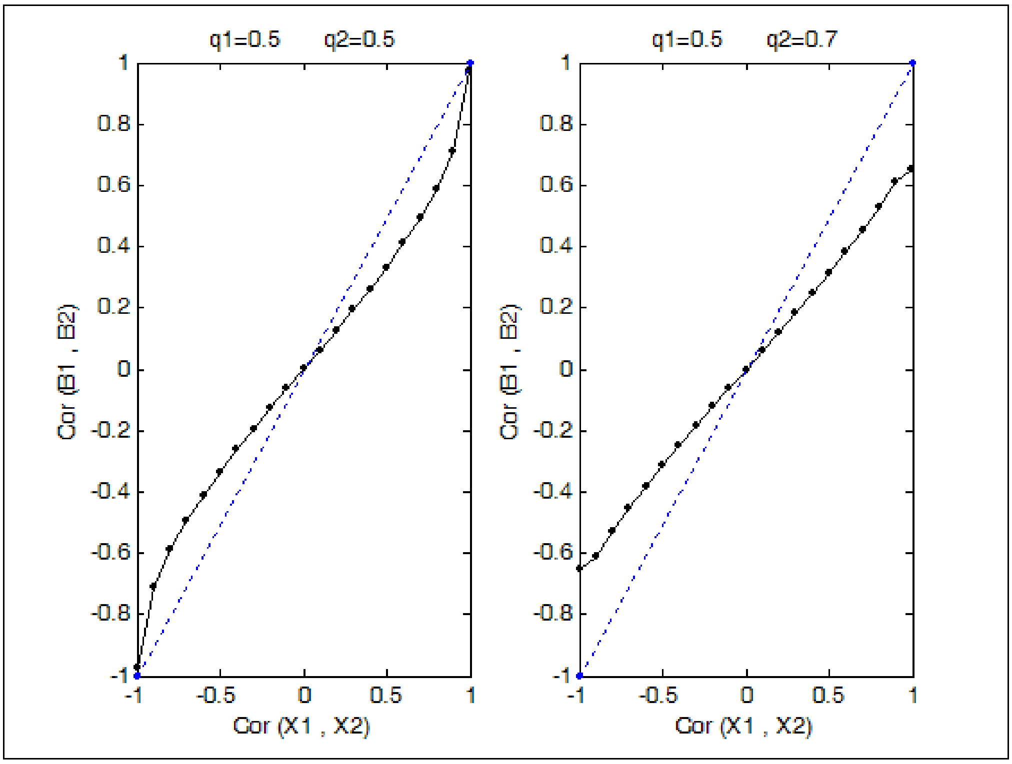
Table 4. $q_1 = 0.4$, $q_2 = 0.6$, $n = 10^7$ ***simulations with MCRCC***

c_{12}	-0.999	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4
r_{12}	-0.9713	-0.7106	-0.5872	-0.4902	-0.4060	-0.3298	-0.2588
c_{12}	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
r_{12}	-0.1912	-0.1261	-0.0628	-0.0004	0.0616	0.1240	0.1869
c_{12}	0.4	0.5	0.6	0.7	0.8	0.9	0.999
r_{12}	0.2512	0.3173	0.3861	0.4589	0.5364	0.6181	0.6667

Table 5. $q_1 = 0.5, q_2 = 0.7$, $n = 10^7$ ***simulations with MCRCC***

c_{12}	-0.999	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4
r_{12}	-0.6546	-0.6091	-0.5293	-0.4529	-0.3809	-0.3125	-0.2472
c_{12}	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
r_{12}	-0.1838	-0.1219	-0.0608	-0.0002	0.0605	0.1214	0.1834
c_{12}	0.4	0.5	0.6	0.7	0.8	0.9	0.999
r_{12}	0.2469	0.3124	0.3808	0.4530	0.5297	0.6091	0.6546

Remark 8. The differences between the correlation values $r_{12} = Cor(B_1, B_2)$ and $c_{12} = Cor(X_1, X_2)$ are sometimes considerable. *Graphic 2* gives us a suggestive illustration of this aspect (compare the differences between the continuous and dotted curves).



Graphic 2. The ratio between the correlation indices r_{12} and c_{12}

Remark 9. We can use successively *Proposition P4.5* and the discretization procedure *DP* to simulate directly samples from a tree type binary systems. See, for example, the one level tree system depicted in *Figure 1, case 1.3.*

6. Concluding remarks

We discussed two algorithms to generate random variates for a binary system $\underline{B} = (B_1, B_2, B_3, \ldots, B_k)$ with k components.

The algorithm *GDRV* uses as inputs all the probabilities $p_{i_1, i_2, i_3, \ldots, i_k}$, $i_j \in \{0, 1\}$, $1 \le j \le k$, which characterize the binary system $\underline{B}$. It is not so easy to apply practically the procedure *GDRV* for systems $\underline{B}$ which have a lot of components. In this case the quantity $2^k - 1$ of the input data for *GDRV* algorithm becomes extremely large.

For this reason is suggested a new other algorithm based on the discretization procedure *DP* to obtain arbitrary observations from $\underline{B}$. This procedure simulate better the real aspects. The correlation structure of a continuous system $\underline{X}$ is inherited by the binary system $\underline{B}$ resulted after a discretization process. The relation between the correlation coefficients $c_{12} = Cor(X_1, X_2)$ and $r_{12} = Cor(B_1, B_2)$ can be determined by applying *MCRCC* algorithm (see also *Graphic 2*).

References

[1] Agresti, A., *An introduction to categorical data analysis*, John Wiley and Sons, New York, 1996.

[2] Andersen, E.B., *Introduction to the statistical analysis of categorical data*, Springer, New York, 1997.

[3] Devroye, L., *Non-uniform random variate generation*, Springer-Verlag, New York, 1986.

[4] James E. Gentle, J.E., *Random number generation and Monte Carlo methods*, Springer - Statistics and Computing, New York, (second edition), 2003.

[5] Leisch, F., Weingessel, A., Hornik, K., *"On the generation of correlated artificial binary data"*, Adaptive Information Systems and Modelling in Economics and Management Science, Working Paper Series SFD, no. 13, Vienna University of Economics, 1998.

[6] Wasserman, S., Faust, K., *Social network analysis: Methods and applications*, Cambridge University Press, New York, 1998.