A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

*Wojciech Gamrot**

ON GENERATING CORRELATED PSEUDO-RANDOM BINARY NUMBERS

Abstract. The paper is devoted to the problem of generating sequences of binary vectors having joint distribution allowing for correlation between individual elements. A procedure for generating such a distribution from uncorrelated binary and multinomial pseudo-random data is proposed. Certain properties of the proposed procedure are examined in the simulation study.

Key words: Random number generator, multinomial distribution, two-point distribution, expectation.

I. INTRODUCTION

Pseudo-random data having multivariate binary (two-point in $\{0,1\}$) distribution find applications in various fields of study. One of such fields is survey sampling where such pseudo-random vectors may be used to simulate the non-controllable stochastic nonresponse mechanism which governs if sampled units provide valid answers in the survey or not. In such a context it is often assumed that events representing the participation or non-participation in the survey are independent. In practice, this assumption does not have to be true, which leads to a growing interest in establishing the properties of estimators under non-independent data missingness. As a result, a need appears to generate pseudo-random multivariate binary data that allow for correlation between individual binary components. This problem has already been considered by many authors including Emrich and Piedmonte (1991), Lee (1993), Gange (1995), Park *et al* (1996) and Leisch *et al* (1998). In this paper another approach to this problem is studied.

II. PROPOSED PROCEDURE

Let $\boldsymbol{x} = [x_1,...,x_k]' \in \{0,1\}^k$ be a random vector with the expected value $E(\boldsymbol{x}) = \mathbf{m}$ where $\mathbf{m} = [m_1,...,m_k]'$ and $E(\boldsymbol{x}\boldsymbol{x}') = \mathbf{M} = [M_{ij}]$. The covariance matrix $V(\boldsymbol{x})=E((\boldsymbol{x}-\mathbf{m})(\boldsymbol{x}-\mathbf{m})')$ of $\boldsymbol{x}$ may be expressed in the form:

* Ph.D., Department of Statistics, University of Economics in Katowice.

$$V(\boldsymbol{x}) = \mathbf{M}-\mathbf{m}\mathbf{m}'$$

For k>2 the knowledge of **M** and **m** does not determine completely the distribution of $\boldsymbol{x}$. Nevertheless, it is desired to generate a sequence of pseudo-random vectors:

$$\mathbf{x}_1,...,\mathbf{x}_n$$

imitating as accurately as possible independent realizations of a random vector $\boldsymbol{x}$ with the expected value **m** and the covariance matrix **M-mm**'. The proposed procedure starts with generating a vector of independent pseudo-random binary variables:

$$\mathbf{g} = [\mathbf{g}_1,...,\mathbf{g}_k]' \in \{0,1\}^k$$

with the expectation

$$E(\mathbf{g}) = \boldsymbol{\gamma} = [\gamma_1,...,\gamma_k]'$$

Then additional independent pseudo-random vectors $\mathbf{f}_1,...,\mathbf{f}_k$ are generated, each of them having multinomial distribution so that

$$\mathbf{f}_i \sim \text{mult}(\boldsymbol{\varphi}_i,k)$$

with $\boldsymbol{\varphi}_1,..., \boldsymbol{\varphi}_k \in <0,1>^k$ being constant vectors of multinomial parameters. Hence, the vectors $\mathbf{f}_1,...,\mathbf{f}_k$ may be arranged in a matrix:

$$\mathbf{F} = [\mathbf{f}_1,...,\mathbf{f}_k]'$$

with the expectation

$$E(\mathbf{F}) = \boldsymbol{\Phi} = [\boldsymbol{\varphi}_1,...,\boldsymbol{\varphi}_k]'$$

where

$$\Phi\ J = J$$

and **J** = [1,...,1]' is the vector of the size k x 1 having each element equal to unity. The binary vector **x** imitating an individual realization of the multivariate random variable $\boldsymbol{x}$ is then obtained according to the formula:

$$\mathbf{x} = \mathbf{F}\mathbf{g}$$

As individual components of **g** are independent, the matrix of their second raw moments may be expressed in the form:

$$E(\mathbf{g}\mathbf{g}') = \boldsymbol{\Gamma}$$

where

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_1 & \gamma_1\gamma_2 & \gamma_1\gamma_3 \\ \gamma_1\gamma_2 & \gamma_2 & \gamma_2\gamma_3 \\ \gamma_1\gamma_3 & \gamma_2\gamma_3 & \gamma_3 \end{bmatrix} = \boldsymbol{\gamma\gamma}' + diag(\boldsymbol{\gamma})(\mathbf{I} - diag(\boldsymbol{\gamma}))$$

with **I** representing an identity matrix of the size k x k. As a result one may express the expectation and the covariance matrix of **x** in the form:

$$E(\mathbf{x}) = h_1(\mathbf{\Phi},\boldsymbol{\gamma})$$

where

$$h_1(\mathbf{\Phi},\boldsymbol{\gamma}) = \mathbf{\Phi}\boldsymbol{\gamma}$$

One may also derive:

$$E(\mathbf{xx}') = \mathbf{\Phi\Gamma\Phi}' - \text{diag}(\mathbf{\Phi\Gamma\Phi}') + \text{diag}(\mathbf{\Phi}\boldsymbol{\gamma})$$

and as a consequence

$$V(\mathbf{x}) = \mathbf{\Phi\Gamma\Phi}' - \text{diag}(\mathbf{\Phi\Gamma\Phi}') + \text{diag}(\mathbf{\Phi}\boldsymbol{\gamma}) - \mathbf{\Phi}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{\Phi}'$$

or equivalently

$$V(\mathbf{x}) = \mathbf{\Phi}(\text{diag}(\boldsymbol{\gamma})(\mathbf{I} - \text{diag}(\boldsymbol{\gamma})))\mathbf{\Phi}' - \text{diag}(\mathbf{\Phi\Gamma\Phi}') + \text{diag}(\mathbf{\Phi}\boldsymbol{\gamma})$$

Assuming **Φ** to be nonsingular, one may express the covariance matrix in the form

$$V(\mathbf{x}) = h_2(\mathbf{\Phi}, \boldsymbol{\gamma})$$

where

$$h_2(\mathbf{\Phi}, \boldsymbol{\gamma}) = \mathbf{\Phi}(\text{diag}(\boldsymbol{\gamma})(\mathbf{I} - \text{diag}(\boldsymbol{\gamma})))\mathbf{\Phi}' - \text{diag}(\mathbf{\Phi}\text{diag}(\boldsymbol{\gamma})(\mathbf{I} - \text{diag}(\boldsymbol{\gamma}))\mathbf{\Phi}') + - \text{diag}(\mathbf{mm}') + \text{diag}(\mathbf{m})$$

so that only two terms in $h_2(.)$ depend on **Φ** and **γ**. Hence, in order to assure that moments of the generated sequence of vectors correctly imitate realizations of ***x***, one may attempt to find the values of **Φ** and **γ** satisfying conditions:

$$\begin{cases} h_1(\mathbf{\Phi},\boldsymbol{\gamma}) = \mathbf{m} \\ h_2(\mathbf{\Phi},\boldsymbol{\gamma}) = \mathbf{M} - \mathbf{mm}' \\ \boldsymbol{\gamma} \in <0,1>^k \\ \mathbf{\Phi} \in <0,1>^{k\cdot k} \\ \mathbf{\Phi J} = \mathbf{J} \end{cases} \tag{1}$$

To eliminate equality constraints associated with elements of the matrix **Φ** one may introduce another matrix $\mathbf{Z} = [z_{ij}]$ of the size k x k and $z_{ij} \in R$ for

i,j=1,...,k. The elements of $\mathbf{\Phi}$ may then be expressed as a transformation of corresponding elements of $\mathbf{Z}$ according to the formula:

$$\varphi_{ij} = \frac{|z_{ij}|}{\sum_{r=1...k} |z_{ir}|}$$

Under assumption that z_{ij} are not simultaneously equal to zero in any row of $\mathbf{Z}$ this assures that all the elements of $\mathbf{\Phi}$ are in the <0,1> range, and their row sums are equal to unity. Moreover, for any possible value of $\mathbf{\Phi}$ there exists some value of $\mathbf{Z}$ that transforms to it. Consequently, instead of finding $\mathbf{\Phi}$ and $\boldsymbol{\gamma}$ satisfying (1) one may attempt to find such $\mathbf{Z}$ and $\boldsymbol{\gamma}$ that:

$$\begin{cases} h_1(\mathbf{\Phi}(\mathbf{Z}),\boldsymbol{\gamma}) = \mathbf{m} \\ h_2(\mathbf{\Phi}(\mathbf{Z}),\boldsymbol{\gamma}) = \mathbf{M} - \mathbf{m}\mathbf{m}' \\ \boldsymbol{\gamma} \in <0,1>^k \\ \mathbf{Z} \in R^{k \cdot k} \end{cases} \tag{2}$$

A solution to this problem may be found by minimizing the following criterion function:

$$Q(\mathbf{Z},\boldsymbol{\gamma}) = \boldsymbol{\delta}\boldsymbol{\delta}' + \mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{\Delta}') + \Psi(\boldsymbol{\gamma})$$

where

$$\boldsymbol{\delta} = h_1(\mathbf{\Phi}(\mathbf{Z}),\boldsymbol{\gamma}) - \mathbf{m}$$

$$\boldsymbol{\Delta} = h_1(\mathbf{\Phi}(\mathbf{Z}),\boldsymbol{\gamma}) - (\mathbf{M} - \mathbf{m}\mathbf{m}')$$

while

$$\Psi(\boldsymbol{\gamma}) = \sum_{i=1,...,k} \max(0, (\gamma_i - 0.5)^2 - 0.25)$$

The first two terms of the function Q are equivalent to the sum of squared differences between desired and actual components of the expectation vector and covariance matrix while $\Psi(\boldsymbol{\gamma})$ is an additional penalty term that forces the elements of the vector $\boldsymbol{\gamma}$ to fall into the <0,1> interval. The minimum possible value for Q is equal to zero. It may be achieved for various combinations of $\boldsymbol{\gamma}$ and $\mathbf{Z}$ values and, if achieved, it guarantees that a solution to the problem (2) is found, although in general Q does not have to be convex. The 'creeping simplex' algorithm of Nelder and Mead (1965) is particularly useful for finding the minimum of Q as it does not require the knowledge of the gradient for Q. The quasi-Newton procedures using finite-difference gradient estimates also seemed to work well in experiments (this was especially true for Broyden-Fletcher-

Goldfarb-Shanno (1970) method implemented in *R*). In any case when zero value is not achieved for Q, it appears reasonable to restart the procedure from other randomly chosen starting point. The failure to achieve zero after carrying out a prescribed number of restarts may be treated as evidence that for given values of **M** and **m** the criterion function does not have a global minimum yielding the value of Q equal to zero. In such a case the proposed procedure may be restarted with increased length of vectors **g** and $\mathbf{f}_1,...,\mathbf{f}_k$.

III. SIMULATION RESULTS

A simulation study was conducted to compare the proposed procedure (in the sequel denoted by the abbreviation: 'PRO') to the well-known algorithm of Leisch et al (1998) and to the procedure of Park et al (1996) – respectively abbreviated by 'LWH' and 'PPS'. The procedures PRO and PPS were implemented in *R* for this study while for LWH an implementation from the R package *bindata* was used. The simulation experiment was carried out for desired parameters of the distribution given by the expectation vector:

$$\mathbf{m} = [0.45, 0.55, 0.65, 0.75]'$$

and the correlation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & 0.35 & 0.30 & 0.20 \\ 0.35 & 1 & 0.25 & 0.15 \\ 0.30 & 0.25 & 1 & 0.10 \\ 0.20 & 0.15 & 0.10 & 1 \end{bmatrix}$$

which corresponds to a desired matrix of second moments:

$$\mathbf{M} = \begin{bmatrix} 0.4500000 & 0.3341250 & 0.3636868 & 0.3805842 \\ 0.3341250 & 0.5500000 & 0.4168223 & 0.4448132 \\ 0.3636868 & 0.4168223 & 0.6500000 & 0.5081534 \\ 0.3805842 & 0.4448132 & 0.5081534 & 0.7500000 \end{bmatrix}$$

A simulation was carried out by generating 10^7 realizations of random vectors with each procedure. Second moments registered during simulation (equivalent to empirical frequencies of joint occurrence of ones for each pair of variables) for each procedure are as follows:

$$W_{PRO} = \begin{bmatrix} 0.4501784 & 0.3342347 & 0.3637991 & 0.3808378 \\ 0.3342347 & 0.5499582 & 0.4168463 & 0.4448244 \\ 0.3637991 & 0.4168463 & 0.6499371 & 0.5082069 \\ 0.3808378 & 0.4448244 & 0.5082069 & 0.7499787 \end{bmatrix}$$

$$W_{LWH} = \begin{bmatrix} 0.4499538 & 0.3339403 & 0.3634486 & 0.3801617 \\ 0.3339403 & 0.5499983 & 0.4175299 & 0.4453388 \\ 0.3634486 & 0.4175299 & 0.6497930 & 0.5082309 \\ 0.3801617 & 0.4453388 & 0.5082309 & 0.7502262 \end{bmatrix}$$

$$W_{PPS} = \begin{bmatrix} 0.4498139 & 0.3339032 & 0.3635074 & 0.3803358 \\ 0.3339032 & 0.5498768 & 0.4168319 & 0.4446168 \\ 0.3635074 & 0.4168319 & 0.6499634 & 0.5080016 \\ 0.3803358 & 0.4446168 & 0.5080016 & 0.7498585 \end{bmatrix}$$

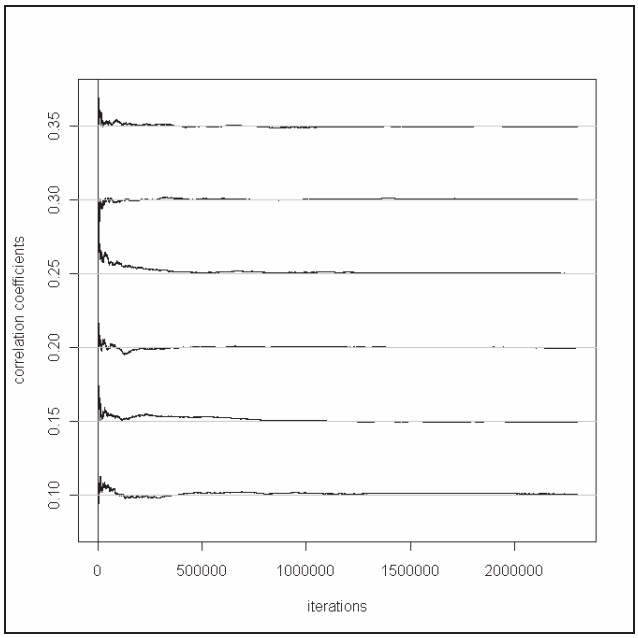
All these frequencies are reported with 7 significant digits so actual counts may be re-computed via multiplying them by 10^7. The null hypothesis stating the equality of respective second moments of generated distributions to desired ones was tested using the Clopper-Pearson (1934) exact test. Two-sided significances (p-values) corresponding to all registered frequencies given above are as follows:

$$\mathbf{P}_{PRO} = \begin{bmatrix} 0.2568009 & 0.4620628 & 0.4605860 & 0.0986609 \\ 0.4620628 & 0.7904716 & 0.8779144 & 0.9431825 \\ 0.4605860 & 0.8779144 & 0.6766616 & 0.7350552 \\ 0.0986609 & 0.9431825 & 0.7350552 & 0.8763852 \end{bmatrix}$$

$$\mathbf{P}_{LWH} = \begin{bmatrix} 0.7692556 & 0.2157395 & 0.1173902 & 0.0059276 \\ 0.2157394 & 0.9913783 & 0.0000056 & 0.0008239 \\ 0.1173901 & 0.0000056 & 0.1699400 & 0.6239795 \\ 0.0059276 & 0.0008239 & 0.6239795 & 0.0986230 \end{bmatrix}$$

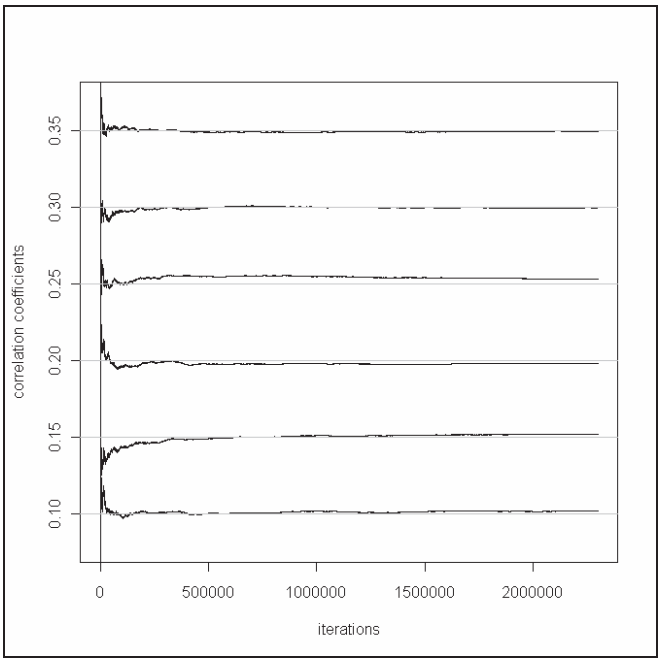
$$\mathbf{P}_{PPS} = \begin{bmatrix} 0.2369632 & 0.1371036 & 0.2382799 & 0.1056975 \\ 0.1371036 & 0.4335622 & 0.9511577 & 0.2114967 \\ 0.2382799 & 0.9511577 & 0.8082722 & 0.3371162 \\ 0.1056975 & 0.2114967 & 0.3371162 & 0.3014308 \end{bmatrix}$$

Hence, extremely low significances were observed only for the LWH procedure which makes corresponding null hypotheses highly questionable for this procedure. The significances for two other procedures were consistent with respective null hypotheses. In addition, the properties of generated pseudo-random sequences are presented graphically, in the form of correlation coefficients between individual variables computed repeatedly for each iteration of the simulation process. Subsequent values of these coefficients are shown as individual points on picture 1 for the PRO, picture 2 for LWH and on picture 3 for the PPS procedure. The first point for each pair of variables represents their correlation coefficient computed for pseudo-random data from iterations 1 to 5, second for iterations 1 to 6 and so on up to the last point representing iterations 1 to $23 \cdot 10^5$. Desired values of correlation coefficients for each pair of variables are shown as horizontal lines.

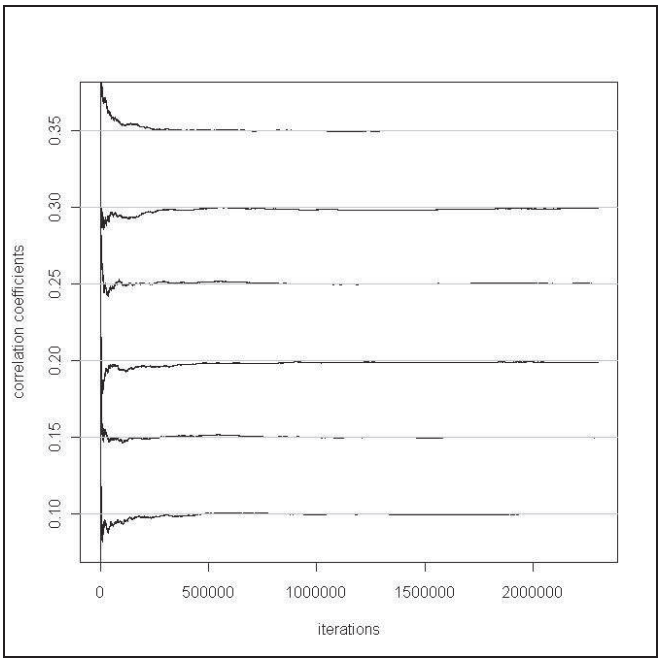


Pic. 1. Correlation coefficients for the PRO procedure

Visual inspection of all three pictures leads to the conclusion that moments of the pseudo-random multivariate data series produced using PRO and PPS procedures are more consistent with specification than in the case of LWH procedure, where subsequent correlation coefficients apparently stabilize at values which differ from desired ones. This supports the results of Clopper-Pearson tests mentioned above.



Pic.2. Correlation coefficients for the LWH procedure.



Pic. 3. Correlation coefficients for the PPS procedure.

IV. CONCLUSIONS

The proposed procedure works in two phases. In the first phase its parameters are calculated numerically. In the second phase correlated pseudo-random binary vectors are produced by transforming uncorrelated binary and multinomial pseudo-random data. Known generators of these distributions are extremely fast, reliable and implemented in most statistical packages. The transformation used in the second phase is extremely simple as well. Hence, it might be expected that the examination of other important properties for the proposed generator (like period length, lack of autocorrelation, reduction of the Marsaglia effect) should also be feasible. The accordance of generated moments with desired ones as shown in this paper constitutes a promising starting point for further analyses. Also, the proposed procedure appears to work quite efficiently when random vectors are to be repeatedly generated for the same desired distribution. During simulation experiments it worked significantly faster than the PPS procedure although not as fast as the LWH. This observation should be interpreted with care as it depends on technical details associated with the implementation as well as operating environment. Anyway, the amount of random access memory used by the proposed procedure is negligible, as opposed to the LWH which allocates large tables.

The range of applications of the proposed procedure is restricted by the fact that it cannot generate negatively correlated binary variables. This limitation might possibly be overcome through merging the vector **g** with its negation. However, such an issue exceeds the scope of this paper.

REFERENCES

Broyden, C. G. (1970) The Convergence of a Class of Double-rank Minimization Algorithms, *Journal of the Institute of Mathematics and Its Applications*, 6, 76-90,

Clopper, C. J., Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413.

Emrich L.J., Piedmonte M.R., (1991) A method for generating high-dimensional multivariate binary variables, *American Statistician*, 45, 302–304.

Gange S.J., (1995) Generating Multivariate Cathegorical Variates Using the Iterative Proportional Fitting Algorithm, *American Statistician*, 49, 134–138.

Lee A.J. (1993) Generating random binary deviates having fixed marginal distributions and specified degrees of association, *American Statistician*, 47, 209–215.

Leisch F., Weingessel A., Hornik K. (1998) On the Generation of Correlated Artificial Binary Data, Working Paper Series SFB, *Adaptive Information Systems and Modelling in Economics and Management Science,* Vienna University of Economics.

Nelder J.A., Mead R. (1965) A simplex method for function minimization, *Computer Journal,* 7, 308–313.

Park C.G., Park T., Shin D.W. (1996) A Simple Method for Generating Correlated Binary Variates, *The American Statistician*, 50(4), 306–310.

Wojciech Gamrot

O GENEROWANIU SKORELOWANYCH BINARNYCH LICZB PSEUDOLOSOWYCH

Niniejszy artykuł poświęcony jest problemowi generowania ciągów wektorów binarnych liczb pseudolosowych dla zadanego wektora prawdopodobieństw brzegowych oraz macierzy korelacji. Zaproponowano procedurę generującą takie wektory na podstawie danych pseudolosowych o rozkładzie zerojedynkowym i wykładniczym. Zbadano wybrane własności zaproponowanej procedury.