ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 123, 1992

*Ene-Margit Tiit**

GENERATION OF MULTIVARIATE RANDOM VECTORS WITH GIVEN CORRELATION MATRIX

1. INTRODUCTION

The generation of random vectors of arbitrary dimension, with given (normal or arbitrary non-normal) marginal distributions and given dependence structure (described by the correlation or covariation matrix), is needed for solving several problems (see A n d e l, 1983):

1) building of imitational models for economic, ecological or social processes;

2) investigating the properties of statistics, especially when the traditional assumption (multivariate normality of initial distribution) is not fulfilled;

3) checking and illustrating theoretical results in multivariate statistics;

4) testing the software for statistical calculation;

Since in most cases one of the following multivariate distributions has been used:

1) multivariate normal or finite mixture of multivariate normals;

2) distribution with independent (or slowly dependent) marginals;

3) distributions belonging to some special families, such as Morgenstern-Farlie or Plackett.

---

* Professor at the University of Tartu, Tartu.

Most practically useable multivariate distributions, including empirical distributions, obtained from experiment, cannot be used in simulation study because there are no effective generators for them.

The aim of the report is to introduce a new method of generating multivariate random vectors with given marginal distributions and given correlation matrix.

The idea of the method is based on two principles:

1) the linear decomposition of correlation matrix in some class of simple correlation matrices.

2) the expression of multivariate distribution in the form of finite mixture of some degenerated distributions.

2. MINIMAL AND MAXIMAL CORRELATIONS FOR THE GIVEN MARGINAL DISTRIBUTIONS

Let P and Q be two univariate distributions with distribution functions F(x) and G(y) correspondingly. Here and later we suppose that for all univariate distributions the second moment exists. Then the extremal bivariate distributions with minimal and maximal correlations are defined by their distribution functions in the following way (see F r e c h e t, 1951, H o e f f d i n g, 1940):

$$H_*(x, y) = \max(0, F(x) + G(y) - 1) \quad (1)$$

$$H^*(x, y) = \min(F(x), G(y)) \quad (2)$$

For all the possible bivariate distributions with marginals F(x) and G(y) the correlation coefficient r fulfills the condition

$$-1 \le r_* \le r \le r^* \le 1 \quad (3)$$

where r_* and r^* are the correlation coefficients calculated by the distributions H_* and H^* correspondingly.

1. In the case when P and Q are continuous, the bivariate distribution T^*, described by (2), is degenerated on the continuous curve on the x / y plane. So as substitution $x \to -x$ allows to reduce the case (1) to (2), we can conclude that distribution T_* has the same property. Figure 1 illustrates the distribution T^* for the case $P = \mu(-c, c)$ and $Q = N(0, \sigma^2)$.

In the case P = Q the supports of the bivariate distributions T^* and T_* have the form of diagonal lines (see Figure 2), if P is symmetrical.

Fig. 1

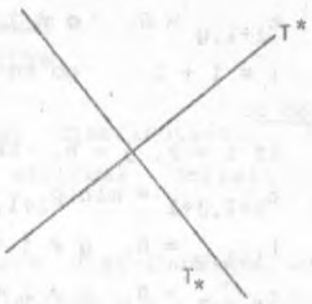

Fig. 2

2. In the case when P and Q have finite support, let $x(P) = k$, $x(Q) = h$, the distributions T^* and T_* have finite support, too.

Let us regard the algorithm of construction of the distribution T^*.

ALGORITHM 1. Our aim is to construct the distribution:

Step 1:

$i = 1, \quad j = 1, \quad t_{11} = \min(p_1, q_1)$.

Step 2:

if $p_i > \sum_{g=1}^{j} t_{ig}$ and $q_j = \sum_{f=1}^{i} t_{fj}$ then step 3;

if $p_i = \sum_{g=1}^{j} t_{ig}$ and $q_j > \sum_{f=1}^{i} t_{fj}$ then step 4;

if $p_i = \sum_{g=1}^{j} t_{ig}$ and $q_j = \sum_{f=1}^{i} t_{fj}$, then step 5.

Step 3:

$$t_{i,j+1} = \min(p_i - \sum_{g=1}^{j} t_{ig}, q_{j+1});$$

$$t_{f,j+1} = 0, \ f \neq i; \quad f = 1, \ldots, k;$$

$j = j + 1$, go to step 2.

Step 4:

$$t_{i+1,j} = \min(q_j - \sum_{f=1}^{i} t_{fj},\ p_{i+1});$$

$t_{i+1,g} = 0, \quad g \neq j, \quad g = 1, \ldots, h;$

$i = i + 1$ go to step 2.

Step 5:

if $i = k$, $j = h$, then end;

$t_{i+1,j+1} = \min(p_{i+1},\ q_{j+1});$

$t_{i+1,g} = 0, \quad g \neq j + 1, \quad g = 1, \ldots, h;$

$t_{f,j+1} = 0, \quad f \neq i + 1, \quad f = 1, \ldots, k;$

$i = i + 1, \quad j = j + 1,$ go to step 2.

To construct the distribution T_*, algorithm 1 can be used for distributions P and H(Q), H(Q) defined with help of the substitution

$$q_i = q_{h-i+1} \quad (i = 1, \ldots, h) \tag{4}$$

The distributions T_* and T^* are monotonous, their supports are well-ordered in the following sense: if (x, y) and (u, v) are two points from support of the distribution T^*, $T^*(x, y) \neq 0$, $T^*(u, v) \neq 0$, then one of the following two relations is true:

$u \leq x$ and $v \leq y$

or

$u \geq x$ and $v \geq y$

E x a m p l e 1. Let P = (0.3, 0.2, 0.2, 0.2, 0.1), Q = = (0.1, 0.4, 0.3, 0.2). We suppose that the supports of the distributions are integers' intervals 1, ..., k and 1, ..., h. The distributions T^*, T_* are given on Figure 3:

y / x	1	2	3	4	5
4				0,1	0,1
3			0,2	0,1	
2	0,2	0,2			
1	0,1				

T^*

y / x	1	2	3	4	5
4	0,2				
3	0,1	0,2			
2			0,2	0,2	
1					0,1

T_*

Fig. 3

For the distributions T_*, T^* we have $r_* = -0.933$ and $r^* = 0.917$.

It is evident that $r^* = 1$ if and only if the distribution vectors P and Q are equal, and $r_* = -1$, if and only if the substitution (4) makes H(Q) equal to P.

3. EXTREMAL CORRELATION MATRICES FOR GIVEN MARGINAL DISTRIBUTION

Let $P_1, P_2, \ldots, P_k$ be given marginal distributions. For every pair of indices i, j there exist extremal correlations r_{*ij}, r^*_{ij}.

Let us describe the set of all k-variate distributions with marginals $P_1, \ldots, P_k$ and extremal correlations r_{*ij}, r^*_{ij}. From this point of view, we define the index-vector I as any subvector of vector $I_o = (1, 2, \ldots, k)$, including the first element 1. Let J be the subvector of I_o complementary to I (that means, $I \cap J = \emptyset$, $I \cup J = I_o$). The number of different index-vectors I equals to 2^{k-1} see T i i t (1984).

For every I there exists k-variate distribution P_I with marginal distributions $P_1, \ldots, P_k$, defined by their distribution functions $F_I(x_1, \ldots, x_k)$ in the following way:

$$F_I(x_1, \ldots, x_k) = \max(0, \min_{i \in I} F_i(x_i) + \min_{j \in J} F_j(x_j) - 1) \qquad (5)$$

The distribution P_I has correlation matrix $R_I = (r^I_{ij})$, defined in the following way:

$$r^I_{ij} = \begin{cases} r^*_{ij}, & \text{if } (i \in I \wedge j \in I) \vee (i \in J \wedge j \in J) \\ r_{*ij}, & \text{if } (i \in I \wedge j \in J) \vee (i \in J \wedge j \in I) \end{cases}$$

Let us denote the set of all matrices R^I by $\mathcal{R}$, $\varkappa(\mathcal{R}) = 2^{k-1}$.

4. LINEAR DECOMPOSITION OF CORRELATION MATRIX IN CLASS

Let R be an arbitrary correlation matrix of range k. We say that R has linear decomposition in class $\mathcal{R}$ if the equation (see A n d e r s o n, 1973):

$$R = \sum_{i=i}^{m} \gamma_i R_{I_i}, \qquad R_{I_i} \in \mathcal{R} \tag{6}$$

holds for

$$\gamma_i \geq 0, \quad \sum_{i=1}^{m} \gamma_i = 1 \tag{7}$$

If the decomposition exists, then $m \leq \frac{k\,(k-1)}{2} + 1$

The necessary (but not sufficient) condition for existance of the decomposition is fulfilling the condition

$$r_{*ij} \leq r_{ij} \leq r^{*}_{ij} \qquad (i, j = 1, \ldots, k) \tag{8}$$

5. THE CONSTRUCTION OF K-VARIATE DISTRIBUTION WITH GIVEN MARGINAL DISTRIBUTIONS AND GIVEN CORRELATION MATRIX

In the case when the decomposition (6) exists for given marginals $P_1, \ldots, P_k$ and correlation matrix R, then we may define the multivariate distribution P in the following way

$$P = \sum_{i=1}^{m} \gamma_i P_{I_i} \tag{9}$$

where P_{I_i} is the distribution having distribution function $F_{I_i}(\cdot)$ (5), see (T i i t, 1984, 1986).

From the properties of finite mixtures, we can draw the following conclusions:

1) all distributions P_I have marginals $P_1, \ldots, P_k$ so their mixture P has the same marginals;

2) the correlation matrix R_p of distribution (9) equals to the following sum

$$R_p = \sum_{i=1}^{m} \gamma_i R_{I_i} = R,$$

consequently, the distribution P has the desired properties.

In general, the distribution P is not unique. The set of all distributions $P = P(P_1, \ldots, P_k, R)$, defined as mixtures of P_{Ii}, has the form of convex polyhedron with vertices, found by formals (9) with different sets $(P_{I_1}, \ldots, P_{I_m})$.

6. THE CONSTRUCTION OF GENERATOR FOR THE RANDOM VECTOR

Let us assume that all distributions P_i are discrete, $\varkappa(P_i) = h_i$.

From the construction given in article (A n d e r s o n 1973) follows that the extremal bivariate distributions are equivalent with some univariate distributions with well-ordered support.

There will be given algorithm 2 for constructing the equivalent univariate distribution $T = (t_f)$ for the bivariate extremal distribution $T^* = (t^*_{ij})$. The distribution T satisfies the following condition: if $t_f = t^*_{ij}$ and $t_g = t^*_{uv}$, then $f \leq g \Leftrightarrow i \leq u$ and $j \leq v$ for $i, u = 1, \ldots, h$; $j, v = 1, \ldots, l$; $f, g = 1, \ldots, w$

ALGORITHM 2. The extremal distribution $T^* = (t^*_{ij})$, $i = 1, \ldots, h$; $j = 1, \ldots, l$ constructed by algorithm 1 is given.

Step 1: $i = 1$, $j = 1$, $f = 1$.

Step 2: $t_f = t^*_{ij}$.

Step 3: if $i = h$ and $j = g$, then $w = f$; end.

Step 4: $i = f + 1$;

if $t_{i+1,j} = 0$, $t_{i,j+1} \neq 0$ then $j = j + 1$, step 2;

if $t_{i,j+1} = 0$, $t_{i+1,j} \neq 0$, then $i = i + 1$, step 2;

if $t_{i+1,j} = 0$, $t_{i,j+1} = 0$, then $i = i + 1$, $j = j + 1$, step 2.

E x a m p l e 2. The distribution T^* calculated in example 1 is equivalent to the following univariate distribution

i	1	2	3	4	5	6	7
vector	(1.1)	(1.2)	(2.2)	(3.3)	(4.3)	(4.4)	(5.4)
probability	0.1	0.2	0.2	0.2	0.1	0.1	0.1

Analogously, the distribution T_* is equivalent to

i	1	2	3	4	5	6
vector	(1.4)	(1.3)	(2.3)	(3.2)	(4.2)	(5.2)
probability	1.2	0.1	0.2	0.2	0.2	0.1

Here the new indices of distribution depend monotonously upon the difference of vector coordinates.

The usage of algorithm 2 for constructing extremal distribution of higher degree (multivariate extremal distributions) will be given.

ALGORITHM 3. The univariate distributions $P_n = (p_1^{(n)}, \ldots, p_{h_n}^{(n)})$, $n = 1, \ldots, k$ and index-vector I are given.

Step 1: $m = 2$; $Q_1 = P_1$; $h = h_1$; $G^{(1)} = (g_j^{(1)})$, $g_j^{(1)} = j$, $j = 1, \ldots, h$.

Step 2: if $m > k$, then step 5;

if $m \leq k$, then $Q_2 = P_m$; $1 = h_m$.

Step 3: if $m \in I$, then $Q_2 = H(Q_2)$, when $H(q_g^{(2)}) = q_{2-g+1}^{(2)}$, $g = 1, \ldots, 1$.

Step 4:

1) by algorithm 1, the extremal bivariate distribution $T^* = (t_{ij}^*)$ from univariate distributions Q_1, Q_2 is constructed, $\varkappa(T^*) = w$;

2) by algorithm 2 the equivalent univariate distribution $T = (t_f)$ of $T^* = (t_{ij}^*)$ is constructed;

3) by the $(m - 1) \times h$ matrix $G^{(m-1)}$ with columns $g_i^{(m-1)}$, $i = 1, \ldots, h$, the $m \times w$ matrix $G^{(m)}$ with columns $g_f^{(m)}$ is constructed in the following way:

$$g_f^{(m)} = g_{\substack{\cdots\cdots \\ j}}^{(m-1)} \quad \text{if} \quad t_f = t_{ij}^*, \quad i = 1, \ldots, h, \quad j = 1, \ldots, 1,$$

$f = 1, \ldots, w$.

4) by $m = m + 1$; $Q_1 = T$; $h = w$, step 2

Step 5: the extremal multivariate distribution is given by the $(k + 1) \times w$ - matrix $\binom{G(k)}{T}$, where the rows $1, \ldots, m$ indicate the values (indices) of components of random vector the last row; end.

E x a m p l e 3. Let P_1 and P_2 be given in example 1, $P_3 = (0.5, 0.5)$; $I = (1, 2, 3)$.

Then the distribution P_I (with all positive dependencies) has the following form:

f	1	2	3	4	5	6	7
$G^{(m)}$	1 1 1	1 2 1	2 2 1	3 3 2	4 3 2	4 4 2	5 4 2
T	0.1	0.2	0.2	0.2	0.1	0.1	0.1

After the equivalent univariate distribution is found, the generator can be constructed in the standard way, due to the fact that in every step some value of a k-variate vector is generated.

7. THE CONSTRUCTION OF GENERATOR FOR THE RANDOM VECTOR WITH GIVEN CORRELATION MATRIX AND GIVEN EQUAL SYMMETRICAL MARGINALS

Let all marginals $P_i = P_o$ $(i = 1, \ldots, k)$ be symmetrical. Then for all i, j we have $r_{*ij} = -1$, $r^*_{ij} = 1$ and, consequently, the class $\mathcal{R}$ of matrices R_I does not depend on a concrete form of distribution P_o.

All distributions P_I are then concentrated on a line directed by any diagonal of unit simplex.

Then every realization of distribution P_I has the following form:

$$\mathcal{Z} = (x_o,\ x_o r^I_{12},\ \ldots,\ x_o r^I_{1k}),$$

where r^I_{1i} $(i = 1, \ldots, k)$ are the elements of the ist row of the correlation matrix R_I, $(x_o \sim P_o)$.

In the case of equal marginal distributions, we get very efficient generator of random vectors with distribution by means of the following algorithm.

ALGORITHM 4.

Step 1: find the linear decomposition of R

$$R = \sum_{i=1}^{m} \gamma_i R_{I_i}, \qquad R_{I_i} \in \mathcal{R}.$$

Step 2: generate the random number α with distribution $\gamma(\gamma_i = 1, \quad i = 1, \ldots, m)$.

Step 3: generate the random number x with distribution P_o and form the vector $\underline{y} = (xr_{1\underline{1}}^{Ii}, \ldots, xr_{1k}^{Ii})$; go to step 2.

It can be concluded, that the generator is rather efficient: to generate one realization of k-variate vector only two random numbers - α and $\underline{y}$ - must be generated.

The same efficiency holds for the case of unequal marginal distributions too, but the preparation (the construction of generators P_I) by algorithm 3 is more labour-consuming.

8. FINAL REMARKS

The methodology presented here has some applications to other regions of statistics as well:

1. The minimal and maximal correlations can be used in data analysis. For a pair of variables, apart from the absolute value of the empirical correlation coefficient, the ratio of the empirical coefficient and the maximal/minimal possible value of it (for a given marginal distributions) is of interest.

2. The linear decomposition of correlation matrices can be regarded as some alternative methodology analogous to factor analysis.

3. The construction of extremal multivariate distributions from a group of (ordered discrete) variables can be regarded as some scaling method.

REFERENCES

A n d e l J. (1983), *Dependent Random Variables with a Given Marginal Distribution*, "Acta Universitatis Carolicusis Mathematicae et Physicae", Vol. 24, No. 1, p. 3-12.

A n d e r s o n T. W. (1973), *Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure*, "Annales of Statistics", Vol. 1, No. 1, p. 135-141.

F r é c h e t M. (1951), *Sur les tableaux de correlation dont les marges sont donnies*, "Annales de Université de Lyon", Sect. A, No. 14, p. 53-77.

H o e f f d i n g W. (1940), *Masstanbinvariate Korrelations theorie*, "Schrifts des Mathematische Institut der Berlin Universität", Bd. 5, p. 179-233.

T i i t E-M. (1984), *The Multivariate Distributions with Given Marginal Distributions and Given Correlation Matrix*, "Tartu Riikliku Ulikoli Toimetised", No. 685, p. 21-36.

T i i t E-M. (1986), *Random Vectors with Given Arbitrary Marginals and Given Correlation Matrix*, "Tartu Riikliku Ulikoli Toimetised", No. 733, p. 14-39.

Ene-Margit Tiit

GENEROWANIE WIELOWYMIAROWYCH WEKTORÓW LOSOWYCH Z DANYMI ROZKŁADAMI BRZEGOWYMI I DANĄ MACIERZĄ KORELACJI

Niech $P_1, P_2, \ldots, P_k$ będą danymi jednowymiarowymi rozkładami z macierzą korelacji R o wymiarach k x k. Powstaje wówczas problem generowania wektorów losowych mających k-wymiarowy rozkład $P(P_1, \ldots, P_k, R)$ z rozkładami brzegowymi P_i i macierzą korelacji R.

Metodą zalecaną do rozwiązania tego problemu jest liniowa dekompozycja macierzy korelacji w klasie macierzy prostych,

$$R = \sum_{l=1}^{m} \gamma_l R_l, \quad R_l \in \mathcal{R}, \qquad \sum_{l=1}^{m} \gamma_l = 1, \quad \gamma_l \geq 0, \quad l = 1, \ldots, m,$$

gdzie $\mathcal{R}$ jest klasą macierzy korelacji mających minimalne i maksymalne możliwe współczynniki korelacji (w sensie Hoeffdinga i Fréchet'a) dla każdej brzegowej pary (P_i, P_j). Rozkład $P(\cdot)$ jest skonstruowany jako dyskretna mieszanina specjalnych zdegenerowanych rozkładów.

W przypadku, kiedy wszystkie dane brzegowe rozkłady są jednakowe i symetryczne, wówczas proste macierze korelacji mają jako elementy 1 i -1.

Zaproponowana konstrukcja rozkładu $P(P_1, \ldots, P_k, R)$ pozwala zbudować wysoce efektywne generatory wektorów losowych. W artykule pokazano praktyczne wykorzystanie wyników badań Monte Carlo dotyczących tych zagadnień.