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 3(302), 2014

*Jacek Stelmach**

ON GENERATING MULTIVARIATE SAMPLES WITH ARCHIMEDEAN COPULAS

Abstract. Archimedean copulas are one of the most known classes of copulas. They allow modeling the dependencies between variables with small number of parameters. This paper presents a method designated to generate multivariate samples of the same distribution like primary sample with Archimedean copulas. Such generator may be used in Monte Carlo investigations to create multivariate samples.

Apart from theoretical considerations there are presented the examples of application of the method. All the calculations were carried out with R 2.15.0 packages.

Keywords: Archimedean copulas, multivariate samples, permutation tests.

I. INTRODUCTION

Rapid development of computer technology caused the statistical methods that require high computing power began to be more widely used. One of them is Monte Carlo simulation that relies on repeating of statistical experiments for random samples. It allows to model complex processes and phenomena, investigate it conditions rarely occur in reality without time-consuming and costly acquisition of real data. The accuracy of this method however depends on the quality of the sample generator. In the case of multivariate inference it may bring difficulties to generate samples with distributions similar to those existing in the investigated process. Frequently the generators already proven with multivariate normal distribution, uniform or t-Student, well defined and programmed are used. Unfortunately, in reality we deal with much more complex dependencies, where the asymmetry, upper and lower tail dependence, multimodality and correlations between variables are observed. Therefore there is a risk that the Monte Carlo simulation carried out on samples whose distributions are not similar to the real may lead to erroneous conclusions.

This paper presents the investigations whose aim was to verify the possibility to use Archimedean copulas in the generation of multivariate samples with the same distribution as primary sample.

* M. Sc., Department of Statistics, University of Economics, Katowice, jacek.stelmach@polwax.pl.

II. ARCHIMEDEAN COPULAS – SHORT DESCRIPTION

Archimedean copulas are the class of copulas that are defined via their generator $\varphi(t)$.

Assuming that $\varphi(t)$ is continuous and strictly decreasing function (copulas generator), its pseudo inversion $\varphi^{-1}(t)$ can be defined as:

$$\varphi^{-1}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0) \\ 0, & 0 \le t \le \infty \end{cases} \tag{1}$$

And then, Archimedean copula class is defined for a vector $u \in [0,1]^d$ according to formula:

$$C^d(u_1, ..., u_d) = \varphi^{-1}(\varphi(u_1) + ... + (u_d)). \tag{2}$$

Depending on a form of $\varphi(t)$ formula, different types of Archimedean copulas with different properties are used. A detailed review of these copulas was presented by Nelsen (2005) and Joe (1997). Discussed investigations were carried out for three, very common copulas: Clayton, Frank and Gumbel. Table 1 presents distribution, generator and range of values of Θ parameter for bivariate copulas.

Table 1. Distribution, generator function and range of parameter for investigated copulas

Copula name	$C(u,v)$	$\varphi(t)$	Range of Θ
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$	$\theta^{-1}(t^{-\theta} - 1)$	$(0,\infty)$
Frank	$-\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$	$\log\left(\frac{e^{-\theta t} - 1)}{e^{-\theta} - 1}\right)$	$(-\infty,\infty)$
Gumbel	$\exp\left(-(\log(u)^{-\theta} + \log(v)^{-\theta}\right)^{\frac{1}{\theta}}$	$(-\log(t))^{\theta}$	$[1,\infty)$

Source: Nelsen (2005).

III. EXPERIMENT DESCRIPTION

The experiment was carried out for six datasets:

1. Dataset 1. Multivariate normal distribution (3 variables) with zero mean vector and variance/covariance matrix as in table 2. 100 observations were generated. This dataset is used as a reference dataset to check proposed statistical methods.

Table 2. Variance/covariance matrix of dataset 1

Variance/covariance matrix		
1	0.5	0.5
0.5	1	0.5
0.5	0.5	1

2. Dataset 2. 50 observations sampled from dataset 1.

3. Dataset 3. Proposed by Friedman (1991) – model of electronic noise, according to Friedman creates great demands for regression methods, described with a formula:

$$y = \sqrt{\left(x_1^2 + \left(x_2 x_2 - \frac{1}{x_2 x_4} \right)^2 \right)} + e \tag{3}$$

where: x_1, x_2, x_3, x_4 – variables with univariate distribution from ranges: $0<x_1<100$; $40\Pi<x_2<560\Pi$; $0<x_3<1$; $1<x_4<11$; e~N(0,9)**.** The dataset contains 4 variables: x_1, x_2, x_3, y and 100 observations.

4. Dataset 4. 50 observations sampled from dataset 3.

5. Dataset 5. 75 empirical observations of 5 variables: chemical parameters of raw material and ready product of petrochemical process. Figure 1 presents the scatterplot of this dataset.

6. Dataset 6. 50 observations sampled from dataset 5.

First there was checked the multivariate independence of the variables with a method proposed by Genest and Remillard (2004) to verify the hypothesis:

$$H_0: C(u_1,\ldots,u_d)=\Pi(u_1,\ldots,u_d) \tag{4}$$

where: $\prod(u_1,...,u_d) = u_1 \times ... \times u_d$ is the copula of independence (all the variables are independent).

This method used empirical copulas $C_n(u)$ defined by pseudoobservations $\hat{U}_i$:

$$\hat{U}_i = \frac{R_i}{n+1},\ i = 1, ..., n, \tag{5}$$

where R_i is a vector of ranged observations. Then the distribution of empirical copulas is:

$$C_n(u) = \frac{1}{n}\sum_{i=1}^{n}(1\left(\hat{U}_i \le u\right),\ u \in [0,1]^d \tag{6}$$

and the test statistics:

$$S = \int_{[0,1]^d} n\left(c_n(u) - \prod_{i=1}^{d} u_i\right)^2 du \tag{7}$$

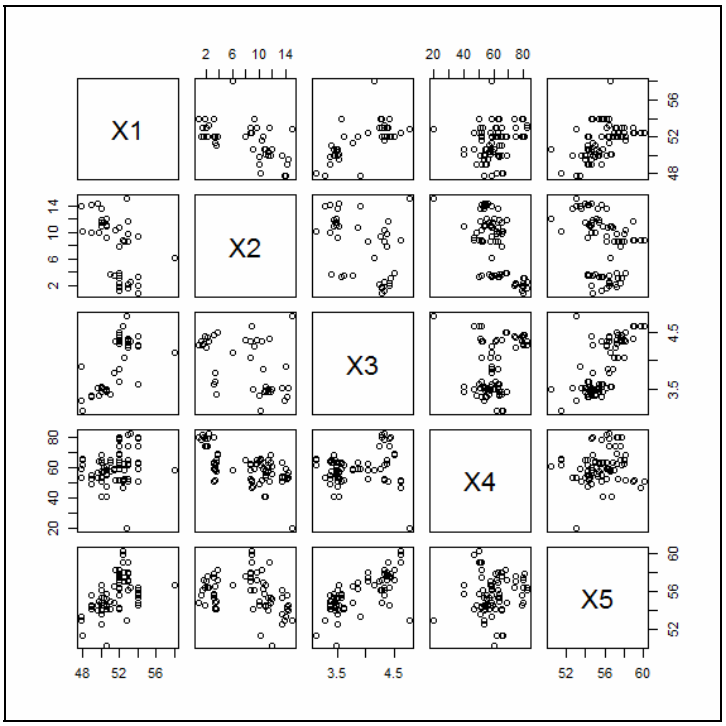

Figure 1 The scatterplot of Dataset 5

To prevent the overlearning during fitting the copulas it was decided to look for the copulas with only one parameter. The estimation of Θ parameter of investigated copulas was carried out maximizing likelihood function:

$$\ln L(\theta; U_{ij}, i = 1, ..., d, \ \ j = 1, ..., n) = \sum_{j=1}^{n} \ln c(U_{1j}, ..., U_{dj}) \tag{8}$$

where:

$$Uij = \frac{1}{n+1} \sum_{k=1}^{n} 1(X_{ik} \le X_{ij})$$

There were carried out 100 iterations with jacknife method (randomly cutting out one observation). It allowed estimating the dispersion of the estimator. Also the mean squared error of fitting was estimated with bootstrap method.

The generating of multivariate samples with fitted copulas was performed with an algorithm proposed by Marshall and Olkin (1988) as a sequence below:

1. Simulate d independent uniform variable u_i for $i = 1,...,d$.
2. Simulate a variable Y with distribution function G such that the Laplace transformation of G is pseudo inversion $\varphi^{-1}(t)$.
3. Wanted multivariate sample is calculated according to formula:

$$X_i = \varphi^{-1}\left(\frac{-\ln(u_i)}{Y}\right) \tag{9}$$

Table 3 contains information about the distribution and density of Y function for investigated copulas.

Table 3. The distribution and the density of Y function

Copula name	Clayton	Gumbel	Frank
Y distribution	Gamma($1/\Theta$, 1)	Stable ($\alpha, \beta, \gamma, \delta$); $\alpha = 1/\Theta, \beta = 1,$ $\gamma = (\cos(\Pi/2\Theta))^{\Theta}$, $\delta = 0$	Logarithmic series for $\alpha=(1-e^{-\Theta})$
Y density		No analytical form.	$P(Y = k) = \frac{-1}{\ln(1-\alpha)} \frac{\alpha^k}{k}$

Source: Marshall and Olkin (1988).

There were carried out 1000 iterations of generation multivariate samples for estimated Θ value. After that the hypothesis about the equality of distribution both: primary sample and generated samples was verified. Although the first two datasets have got normal distribution, the permutation tests were used to compare the results for all the datasets (next dataset have got no normal distribution). There were chosen 3 test statistics presented in table 4.

Table 4. Test statistics used in the experiment to verify the equality of distributions

Test	Test name	Test statistics
1	Permutation test based on T^2– Hotelling test.	$ST_1 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T S_{pl}^{-1}(\bar{x}_1 - \bar{x}_2)$
2	Permutation test based on volume of confidence ellipsoid (see Kończak and Stelmach (2013)).	$ST_2 = \left(\frac{\max(v(1), v(2), v(1 \cup 2))}{\min(v(1), v(2), v(1 \cup 2))}\right) - 1$
3	Permutation test sensitive on difference of means and variance/covariance matrices.	$M = 1 - \frac{\lvert S_1\rvert^{\frac{n_1-1}{2}} \bullet \lvert S_2\rvert^{\frac{n_2-1}{2}}}{\lvert S_{pl}\rvert^{\frac{n_1+n_2-2}{2}}}$; $S_{pl} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1 + n_2 - 2}$ $ST_3 = \left(\frac{T^2}{F_{\alpha,k,n_1+n_2-k-1}}\right)^2 + M^2$

Source: Rencher (2002), Kończak and Stelmach (2013).

IV. EXPERIMENT RESULTS

For all the datasets, the hypothesis concerning the multivariate independence was rejected. It means that copula of independence $\Pi(u_1,\ldots,u_d)$ cannot be used in the experiment.

For fifth and sixth dataset, the estimation of Θ parameter was possible only for Clayton copulas due to numerical limitations. Means and standard deviation of Θ are presented in table 5. Mean square error of fitting the copulas is put in table 6.

Table 5. Mean and standard deviation of estimated Θ parameter

	Mean of Θ			Standard deviation of Θ		
Dataset	Clayton	Frank	Gumbel	Clayton	Frank	Gumbel
First	1.124	4.078	1.617	0.016	0.059	0.011
Second	0.929	3.096	1.450	0.041	0.117	0.019
Third	0.302	1.393	1.181	0.006	0.025	0.004
Forth	0.393	1.714	1.243	0.016	0.074	0.010
Fifth	0.051	–	–	0.004	–	–
Sixth	0.045	–	–	0.005	–	–

Table 6. Mean square error of fitting the copulas

Dataset	Clayton	Frank	Gumbel
First	0.164	0.537	0.104
Second	0.240	0.599	0.096
Third	0.063	0.329	0.038
Forth	0.101	0.518	0.083
Fifth	0.052	–	–
Sixth	0.071	–	–

Mean square error depends on copulas type and dataset and is higher if primary sample has got lower size.

The generation of samples (1000 iterations) with Marshall and Olkin algorithm was possible. The results of permutation tests that verified the equality of distribution of primary sample and generated samples are presented in table 7. For fifth and sixth dataset – that represents real data, two significance levels α=0.05 and α=0.10 were chosen. The results – as a percentage of rejections are placed in table 7.

Table 7. The percentage of rejection the hypothesis about the equality of distributions – primary and generated samples

Copula name	Clayton			Frank			Gumbel		
Test	T1	T2	T3	T1	T2	T3	T1	T2	T3
Dataset 1, α = 0.05	0.000	0.212	0.000	0.000	0.138	0.000	0.000	0.136	0.000
Dataset 2, α = 0.05	0.000	0.03	0.000	0.000	0.041	0.000	0.000	0.002	0.000
Dataset 3, α = 0.05	0.000	0.414	0.000	0.000	0.396	0.000	0.000	0.279	0.000
Dataset 4, α = 0.05	0.000	0.923	0.001	0.000	0.679	0.000	0.000	0.851	0.000
Dataset 5, α = 0.05	0.000	0.634	0.000						
Dataset 6, α = 0.05	0.000	0.615	0.000						
Dataset 5, α = 0.10	0.000	0.395	0.000						
Dataset 6, α = 0.10	0.000	0.586	0.000						

The most interesting results were produced by test no 2. First and third test statistics did not rejected the hypothesis, while the second differentiated the results of experiment depends on copulas type and the distribution (and size) of dataset. Although estimated MSE testified clearly worse fitting for Frank copulas, the results of permutation tests did not confirm it. The most important is the distribution of the samples – the least number of rejections was observed for multivariate normal distribution samples.

V. CONCLUSIONS

1. It is possible to fit the Archimedean copulas to given sample, type of copulas for which the goodness-of-fit is the best depends on the distribution of the sample.

2. Checking the suitability of estimated copulas only with mean square error can be improper. Additional tests that verify the equality of the distributions of primary sample and generated samples are recommended.

3. It is possible to use the estimated copulas to generate the samples for Monte Carlo method purposes. It allows carrying out statistical inference with the type of simulated multivariate samples that closer represents investigated process or phenomenon.

REFERENCES

Fermanian, J.-D. (2005), Goodness of Fit Tests for Copulas, Journal of Multivariate Analysis, 2005, p. 119–152.

Friedman J (1991), Multivariate adaptive regression splines, „Annals of Statistics" 1991, vol 19. Institute of Mathematical Statistics, Stanford University.

Genest Ch., Favre A.-C. (2007), Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask*, Journal of Hydrologic Engineering*, Jul/Aug 2007.

Genest Ch., Remillard B. (2004), Tests of Independence and Randomness Based on the Empirical Copula Process, *Test 13(2)*, p. 335–369.

Genest Ch., Remillard B. (2008), Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models*, Annales de l'Institut Henri Poincare: Probabilites et Statistizues*, 44, p. 1096–1127.

Genest Ch., Remillard B., Beaudoin D. (2007), Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics* 44, October 2009, p. 199–213.

Hofert M. (2007), Sampling Archimedean copulas, *Fakultat fur Mathematik und Wirtschaftswissenschaften*, Universitat Ulm.

Joe H. (1997), *Multivariate Models and Dependence Concepts*, Chapman&Hall/CRC, USA.

Kończak G., Stelmach J. (2013), O porównaniu dwóch populacji wielowymiarowych z wykorzystaniem objętości elipsoid ufności, *Zeszyty Naukowe Wydziałowe 133*, Uniwersytet Ekonomiczny w Katowicach.

Marshall A., Olkin W. (1988), Families of multivariate distributions*, Journal of the American Statistical Association*, 83, p. 834–841.

Nelsen R. B. (2005), *An Introduction to Copulas*, Springer Series in Statistics, USA.

Rota G. C. (1964), On the Foundations of Combinatorial Theory. Theory of Möbius Functions*, Zeitschrift für Wahrscheinlichkeitstheorie und werwandte Gebiete*, 2, p. 340–368.

Savu C., Trede M. (2004), Goodness-of-fit tests for parametric families of Archimedean copulas, *Institute for Econometrics,* University of Munster.

Sklar A. (1959), Fonctions de repartition a n dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Universite de Paris* 8, p. 229–231.

Smith M. D. (2003), Modelling sample selection using Archimedean copulas, *Econometrics Journal 2003*, vol. 6, p. 99–123.

Zimmer D. M., Trivedi P. K. (2005), Copula Modelling: An Introduction for Practitioners, *Foundations and Trends in Econometrics*, vol. 1, no. 1, p. 1–111.

Zimmer D. M., Trivedi P. K. (2006), Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand, *Journal of Business & Economic Statistics*, January 2006, vol. 24, no. 1.

Jacek Stelmach

O GENEROWANIU PRÓB WIELOWYMIAROWYCH ZA POMOCĄ KOPUL ARCHIMEDESA

Kopule Archimedesa należą do najbardziej znanych klas wśród kopul. Pozwalają one na modelowanie zależności pomiędzy zmiennymi za pomocą małej ilości parametrów. Artykuł przedstawia metody, które mogą być wykorzystane do generowania prób wielowymiarowych o rozkładzie takim samym, jak próba pierwotna, wykorzystując kopule Archimedesa. Taki generator może być wykorzystany w badaniach Monte Carlo do tworzenia prób wielowymiarowych. Oprócz teoretycznych rozważań zaprezentowano przykłady zastosowania tych metod. Wszystkie obliczenia wykonano wykorzystując procedury programu R 2.15.0.