MODELING DEPENDENCIES IN BIVARIATE DISTRIBUTIONS

Constantin TÂRCOLEA[1], Adrian Stere PARIS[2], Dana SYLVAN[3]

The study of univariate characteristics for experimental data is straightforward, but it is important to reveal the common description of all attributes and their dependencies. It is always possible to obtain the univariate components from discrete or continuous random vectors. Conversely, in general case, the problem is a complicated one. The target of this paper is to present models for an overall estimation of all the properties and their relationships. Some bivariate models were processed, and explained on the basis of an experimental data set. The correlation matrix of values of the cumulative distribution functions (CDF), obtained with chosen methods, and improves the reliability of the proposed output methods. A univariate Weibull distribution for the bivariate copula real values for two attributes it was estimated as an authentication. The ideas developed in the paper remain valuable for interconnected multivariate cases, described in high-dimensional space, without essential modifications.

Keywords: bivariate analysis, copula, multivariate fitting distributions

1. Introduction and Preliminaries

This paper is limited to the field of applied mathematical physics problems. It is always possible to calculate each distribution of the components, given a multivariate distribution, but conversely it is really possible in the particular cases, for example for the independence of marginal distributions. The copula theory offers opportunities for advanced design and should model correlated structures with well-known characteristics and theirs links; copula functions describe time varying with linear / nonlinear features, marginal distributions, jointly with theirs statistical dependencies [1]. Copula calculi require only marginal *CDF*'s of properties and theirs correlation parameters, in order to approximate the shared model. It is essential in decision support to obtain a reliable assessment, properly accounting and modeling these features, dependencies and correlations. It was proven that the copula techniques are advanced tools for modeling dependence and interdependence structures [2, 3]. The present paper evaluates, for example, the relationship between the hardness and tensile strength properties of steels, useful in

[1] Prof., Dept. of Math., UPB, Romania, e-mail: constantin_tarcolea@yahoo.com
[2] Assoc. Prof., Dept. of CCII, UPB, Romania, e-mail: adrian.paris@upb.ro
[3] Prof., Dept. of Statistics, City University of New York, e-mail: dana.sylvan@gmail.com

practical applications [4]. Some attempts have been made to establish a relationship between these properties for various single structurated steels [5].

It is preferable to measure the hardness of the materials, because it allows an easier estimation, even for a small scale production [6]. The hardness–strength entity describes and evaluates the performance of mechanical properties. For illustration a sample of 25 paired data, Brinell hardness (*H*) and tensile strength (*S*), is given by Sultan [7] in the table 1 from [9].

` *Table 1*

Paired values of (H) and (S)

Prop.	Values								
H	143	200	160	181	148	178	162	215	161
S	34.2	57	47.5	53.4	47.8	51.5	45.9	59.1	48.4
H	141	175	187	187	186	172	182	177	204
S	47.3	57.3	58.5	58.2	57	49.4	57.2	50.6	55.1
H	178	196	160	183	179	194	181		
S	50.9	57.9	45.5	53.9	51.2	57.5	55.6		

The rank correlation coefficients Pearson's rho, Kendall's tau and Spearman's rho were used to evaluate the interdependences of chosen characteristics. Those coefficients are important tools in describing linear links; quantify dependence, and therefore the modeling of copulas, but the non-linear relationship structures are described only by the last two coefficients [8].

Chan analyzed a bivariate process based on these data, and proposed a multivariate process capability index over a general tolerance domain, a generalization of the rectangular and ellipsoidal areas [9].

The materials properties depend on many factors and therefore normal distributions are adequate, at least at the beginning. Pavlina and Tyne [10] compiled values of hardness and tensile strength for some mild steels. Tensile strength of the steels usually exhibits a linear correlation with the hardness over the entire range of strength values. The paper [10] proposes some bivariate copula, which describes the general behavior of some given characteristics.

The goodness-of-fit of the empirical univariate data sets, h_{norm} and s_{norm}, (table 1) to adequate the normal distribution is tested in the first step.

The p-values, calculated with adequate statistical tests, (table 2), proved that the normal distribution should model the experimental univariate values for each subset of data sets. The resulted p-values of statistics for normality fitting should be greater than α, (α=0.05, the chosen significant level), what in these cases was fulfilled for all applied tests (table 2).It should be concluded that the random vector approach gives a global idea for all characteristics, and also about all the interdependences of the properties.

The bivariate normality for joint data was checked in the next step. The following values resulted in this case:
a. Mardia's test [4] checked if a bidimensional set of data approach a bivariate data (Fig.1).

Table 2

Results of univariate normal distribution fitting

Test	Kolmogorov-Smirnov	Anderson-Darling	Lilliefors-van Soest	Cramer-von Mises	Ryan-Joiner
h_{norm}	P=0.483	P>0.2	P=0.32	P>0.1	P=0.792
s_{norm}	P> 0.15	P>0.25	P>0.20	P=0.29	P>0.1

The numerical results are: g_1p = 0.1814531; chi.skewness = 0.7258123; p.value.skewness = 0.948108; g_2p = 6.572155; z.kurtosis = -0.8743728; p.value.kurtosis = 0.3819153; chi.small.skewness = 0.8875422; p.value.small = 0.9263412. The p-values of skewness and kurtosis statistics should be greater than α, (α=0.05, the chosen significant level), for multivariate normality, what in these cases was proved.

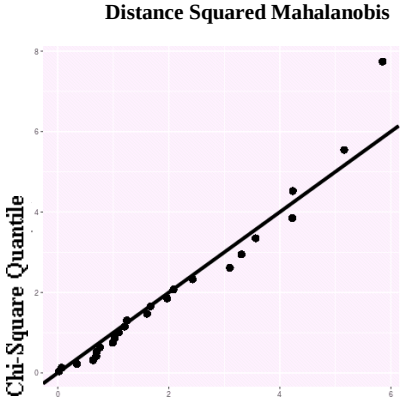


Fig. 1. Chi-Square Q-Q Plot

b. the Henze-Zirkler and Royston tests gave similar results, what indeed justified the bivariate normal distribution as an adequate model.

2. Main results

2.1Nonparametric method for copula calculus

The empirical univariate distribution function *(EUDF)*, as marginal distribution, is given by:

$$EUDF\ (x)\ = \frac{1}{n+1}\sum_{i=1}^{n} I\ \ (x_i \le x) \tag{1}$$

where *I* is the indicator function:

$$I\ (x_i \le x) = 1,\ and,\ I\ (x_i > x) = 0. \tag{2}$$

It is known that the *EUDF* estimates the cumulative distribution, and, if the sample size is sufficiently large, it converges to the theoretical cumulative distribution function of the population [11]. The distributions of margins can be arbitrary; here the values of the empirical distributions of the bivariate data set (table 1) were calculated.

The random variable *H* (hardness) (table 1) has an average $\bar{x}$= 178.625, and an empirical dispersion s^2 = 17.31252; the second random variable *S* (strength), has respectively $\bar{x}$= 53.07083, s^2 = 4.4970022; the value of the correlation coefficient is $\rho\ (H, S)$ = 0.831.

The steps of the method are further in brief described. Consider a random sample, by the size n, from the bivariate population:

h_1, s_1; ...; h_i, s_i; ...; h_n, s_n.

Next were rescaled the observed data in ascending order, namely:

$$\hat{u}_{(i)} = \frac{Rank(h_i)}{n+1}\ ;\quad \hat{v}_{(i)} = \frac{Rank(s_i)}{n+1}\ , i = 1,2, \dots n. \tag{3}$$

The vector:

$$(\boldsymbol{f}, \boldsymbol{g}) = (\hat{u}_{(1)}\ , \hat{v}_{(1)}\ ; \dots;\ \hat{u}_{(i)}\ ,\ \hat{v}_{(i)}\ ; \dots;\ \hat{u}_{(n)}\ , \hat{v}_{(n)}\ )$$

can be considered a random sample for a bivariate copula.

It is known that a cumulative distribution function of the sample of the normalized ranks is a consistent estimation of the true cumulative distribution function [12].

The empirical multivariate distribution function, *EMDF* (copula), is defined as:

$$EMDF(\frac{i}{n}, \frac{j}{n}) = \frac{Card\{(u,v)/\ u \le \hat{u}_{(i)}, v \le \hat{v}_{(j)}}{n} \tag{4}$$

where $\hat{u}_{(i)}$, $\hat{v}_{(j)}$, $1 \leq i, j \leq n$ are the order statistics of the random sample.

The two-dimensional copula *EMDF (F, G)* is a bivariate distribution, with uniformly distributed margins on canonical interval [0; 1]. The one- dimensional values of the bivariate empirical copula where modeled as a univariate *CDF* Weibull (Fig.2), useful in the reliability theory [12].

The resulted Weibull *CDF* has the following expression:

$$F(x) = 1 - \exp(-0.0125x^{1.53}). \quad (5)$$

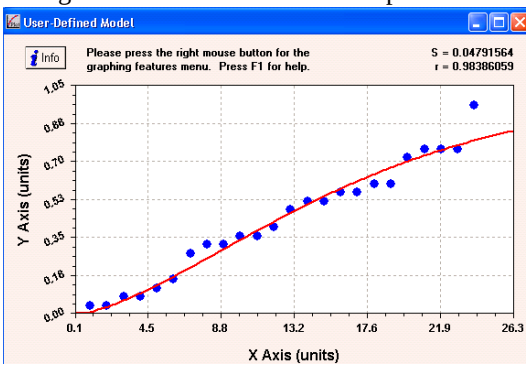
It results that the shape parameter is over unity, what implies that the hazard rate is an ascending one.

2.2 Nataf model for bivariate normal copula

The calculus of a bivariate cumulative normal distribution values is solved first with the Nataf model [3]. The Gaussian copula is a link between a multivariate normal distribution and marginal normal distributions. The bidimensional normal copula, with the given correlation coefficient and characteristics of the components, has the following form:

$$C(u, v) = \Phi(\Phi^{-1}(u), \Phi^{-1}(v)); 0 \leq u, v \leq 1. \quad (6)$$

Fig.2. Weibull estimate of the copula values



Then the joint distribution of those variables resulted from those probabilities using their individual inverse distribution functions.

The Nataf transformation starts with the calculus of $\Phi(x)$ and $\Phi(y)$, and then transfers the range of the formers Φ values into the standard normal variables.

The next step of the transformation is to estimate the Pearson correlation coefficient, here ρ=0.831 [14].

The bivariate cumulative distribution function *G(x, y),* for arbitrary marginal distributions, induced by the Gaussian copula, is:

$$G(x, y) = \Phi 2\ (\Phi^{-1}(F(x)),\ \Phi^{-1}(F(y)), \rho). \quad (7)$$

The values of the bivariate *CDF* should be compiled with the bivariate normal copula function. For example, if *x*=200, *y*=57, then:

F (200) = 0.89152; F(57) =0.80887; Φ^{-1}(0.89152) =1.23466;

Φ^{-1} (0.80887) =0.87373;

it results:

G (200, 57) = Φ_2 (1.23466; 0.87373; 0.831) = 0.78725.

2.3 The method of transforming the dependent into independent variables

The random variables $H \sim N$ (178. 62, 17.31^2) and $S \sim N$ (53 .07, 4.497^2) are correlated with ρ=0.831, and the resultant bivariate normal density function is:

$$f(h,s) = \frac{1}{2\pi\ 17.31*4.49\sqrt{1-0.83^2}} \exp[-\frac{z(h,s)}{2(1-0.83^2)}]$$

where: (8)

$$z(h,s) = \frac{(h-178.62)^2}{17.31^2} - \frac{2*0.83(h-178.62)(s-53.07)}{17.31*4.49} + \frac{(s-53.07)^2}{4.49^2}.$$

It is known that the random variables X, Y, [14], defined by the formula:

$$\begin{cases} X = \frac{H-\mu_H}{\sigma_H} \\ Y = -\frac{\rho}{\sqrt{1-\rho^2}}\ \frac{H-\mu_H}{\sigma_H} + \frac{1}{\sqrt{1-\rho^2}}\ \frac{S-\mu_S}{\sigma_S} \end{cases} \quad (9)$$

are independent and each standard normal distributed. In the studied case it follows:

$$\begin{cases} X = \frac{H-178.62}{17.31} \\ Y = -\frac{0.831}{\sqrt{1-0.831^2}}\ \frac{H-178.62}{17.31} + \frac{1}{\sqrt{1-0.831^2}}\frac{S-53.07}{4.497} \end{cases} \quad (10)$$

The random vector (*X, Y*) has the cumulative distribution function *F(x, y)* equal with the product of the cumulative distribution functions of the components, because they become independent:

$\rho(X,Y)$=-4.08972E-06≈0.

Next there were calculated the values of the cumulative distribution function of the standardized normal random variable *X*, denoting *Φ (x)*, were further calculated, respectively the values of the cumulative distribution function, of the random variable *Y*, denoted *Φ(y)*.

The bivariate cumulative function *F(x, y)* is the product of these independent univariate cumulative distributive functions

$$\boldsymbol{F(x,y) = \emptyset(x) * \emptyset(y)} \quad (11)$$

The correlation coefficients (table 3) were computed to prove the concordance between these three presented models. It is obvious that all three models show a good association; the best correlation is between Nataf and Copula *(EMDF)* models.

Table 3

Correlation matrix of the *CDF*'s values

	G(x,y)	*F(x,y)*	*EMDF(u,v)*
G(x,y)	1		
F(x,y)	0.761104	1	
EMDF(u,v)	0.964079	0.790292	1

3. Conclusions

The bivariate joint distribution of the studied case was written in terms of univariate marginal distribution functions of the exploratory variables and a copula, which models the dependence structure between the predictor variables.

The paper offers an example of calculus of a bivariate statistical distribution for a simultaneous explanation based on two basic mechanical properties, Brinell hardness and tensile strength of steels. The idea should be similarly applied to high-dimensional models, and offers estimates of a joint distribution of many characteristics by evaluating all properties together with their dependences.

Additionally it was proposed a univariate Weibull model for the bivariate copula, given real values for *CDF*, an adequate description of the reliability of the mechanical products, if the attributes are measured.

Mechanical devices are frequently subject to wear and their reliability is closed to normal distributions, because their life depends on many factors.On the other hand the variability of manufactured parts and the quality control procedures are usually modeled by the normal distribution. The introduced nonparametric method for copula calculi solved the problem in the general case, for an unknown multivariate distribution. The calculus of the bivariate cumulative normal distribution was developed with the Nataf model and the method of transforming the dependent into independent variables, with comparable results. The authors proposed a new approach for the analysis of

joint properties, with univariate distributions, using only real numbers, the values of *CDF* of the bivariate functions.

The copula approach allows reliable estimation of multiple correlations for example between the steel alloys components and characteristics, on the basis of the performed studies and analyses of the obtained results. In the future we try to research criteria for hierarchical multivariate distribution functions, based on the marginal distributions data.

R E F E R E N C E S

[1] *A.Paris, D.Sylvan and C.Târcolea,* Maintenance Dependence Modeling with Gaussian Copulas, - Conferinta Internationala "Calitate si Siguranta in Functionare" CCF 2016, Asigurarea Calitatii – Quality Assurance, Anul XXII, nr. 87, 2016, pp. 28-32.

[2] *V. Najjari, H.Bal and S.Celebioglu*, Modeling of dependence structures in meteorological data via archimedean copulas, U.P.B. Sci. Bull., Series D, Vol. 75, Iss. 3, 2013, pp.131-138.

[3] *R.B.Nelsen*, An Introduction to Copulas, Springer Series in Statistics, 2006.

[4] *I.Brooks, P.Lin, G.Palumbo, G.D. Hibbard and U.Erb,* Analysis of hardness–tensile strength relationships for electroformed nanocrystalline materials, Materials Science and Engineering, A, vol. 491, Issues 1–2, (15 september 2008) pp. 412-419.

[5] *M.Umemoto, Z.G.Liu, K.Tsuchiya, S.Sugimoto and M.M.A. Bepari,* Relationship between hardness and tensile properties in various single structured steels, Journal Materials Science and Technology, Volume 17, Issue 5, 2001, pp.505-511.

[6] *P. Zhang, S.X. Li, Z.F. Zhang*, General relationship between strength and hardness, Materials Science and Engineering, A-529, 2011, pp.62–73.

[7] *T.I.Sultan*, An acceptance chart for raw materials of two correlated properties, Quality Assurance, 12(3), 1986, pp. 70-72.

[8] *S.Popova, P.Koprinkova-Hristova, P.Zlateva and A.Toncheva,* Multivariate Analysis of Steel Alloys Components and Characteristics Using Copula Approach, Proceedings of the 3rd International Conference on Application of Information and Communication Technology and Statistics in Economy and Education, Sofia, Bulgaria, 2013, pp.706-713.

[9] *L.K.Chan, S.W. Cheng and F.A. Spiring*, A multivariate measure of process capability, International Journal of Modeling and Simulation, vol.11, no.1, 1991, pp.1-6.

[10] *E. J.Pavlina, C. J.Tyne*, Correlation of Yield Strength and Tensile Strength with Hardness for Steels, Journal of Materials Engineering & Performance. Vol. 17 Issue 6, 2008, pp. 888- 893.

[11] *P.Deheuvels*, La fonction de dépendance empirique et ses propriétés. Un test non parametrique d'independance. Acad. Roy. Belg. Bull. Cl. Sci. (5), 65(6), 1979, pp. 274–292.

[12] *C.Târcolea, A.S. Paris*, Bivariate Weibull Distributions applied to Maintenance Modeling, Asigurarea Calitatii - Quality Assurance, Vol. XXIV, Issue 95, 2018, pp. 16-18.

[13] *K. V.Mardia*, Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 1970, pp. 519–530.

[14] *A. S. Paris, C.Târcolea*, Modeling of Bivariate Data for Dependability, 11th Symposium Durability and Reliability of Mechanical Systems, SYMECH 2018, Fiability and Durability, Ed.Acad., Tg. Jiu, no. 1(21)/2018, pp. 257-260.