INFERENCES IN A COPULA MODEL FOR BIVARIATE SURVIVAL DATA

Mariana CRAIU, Corina CIPU, Laura PÂNZAR*

Obiectivul acestui articol este de a estima parametrii unui model bidimensional de defectare pe baza unei selectii aleatoare $(T_{1j}, T_{2j})_{j=\overline{1,n}}$ *folosind metode parametrice si semiparametrice. Asocierea celor doua variabile* T_1 *si* T_2 *se modeleaza prin copule si se compara rezultatele studiului de inferenta statistica. Copula reprezinta un mod natural de masura a dependentei dintre variabilele aleatoare.*

The aim of this paper is to estimate the parameters in a bivariate lifetime model in the light of a random sample $(T_{1j}, T_{2j})_{j=\overline{1,n}}$ *by parametric or semi-parametric methods. We model the association of the bivariate failure times by copulas, and compare the results of statistical inference. Copulas provide a natural way to study and measure dependence between random variables.*

Keywords: Weibull distribution, bivariate distribution, copula's family, dependence parameters, semiparametric methods.
Mathematics Subject Classification 2000: 62H12.

Introduction

In many cases it is convenient to express a joint distribution $F(x, y)$ as a function of $F_X(x)$ and $F_Y(y)$ (the individual distribution functions for variables X, Y) by:

$$F(x, y) = C(F_X(x), F_Y(y)) = F(F_{X^{-1}}(F_X(x)), F_{Y^{-1}}(F_Y(y))).$$

In this way the mapping C (that is uniquely determined on the unit square when F_X and F_Y are continuous) captures the dependence between the random variables X and Y.

In the last years many research papers develop multivariate survival distributions. A multivariate distribution is derived assuming that marginal distributions are of some specified family. In studies of reliability components are assumed to have independent lifetimes but, is more realistic to assume that there exist some sort of dependence among components.

* Prof., lecturer, assist., Dept. of Mathematics III, University "Politehnica" of Bucharest, ROMANIA

A useful way to develop bivariate lifetime models is through a family of copulas $C(u, v, \delta)$ with a specification of the marginal distributions (where δ is a parameter that determines the dependence structure).

A bivariate copula $C(u, v, \delta)$ is a family of distribution functions (with u, v uniform marginals) defined in $[0,1]^2$ with C(u,1)=u, C(1,v)=v, C(u,0)=C(0,v)=0.

For any copula C there exist two copulas: the Frechet-Hoeffding upper bound defined by M(u,v)=min(u,v) (represents the most positive dependence with each variable being an increasing monotone transformation of any other variable) and the Frechet lower bound $W(u, v) = \max\{0, u + v - 1\}$ (this represents the most negative dependence when one variable is a decreasing monotone transformation of the other variable) for which:

$$W(u, v) \le C(u, v) \le M(u, v).$$

By the Sklar's Theorem [6], for any joint distribution function F with marginals F_1 and F_2, there is a copula C such that for all real numbers x, y:

$$F(x, y) = C(F_1(x), F_2(y)).$$

And conversely, if C is a copula and F_1, F_2 are univariate functions, then the function F(x, y) is a joint distribution with marginals F_1 and F_2.

Between different families of copulas, a special class is that of **Archimedean copulas**.

- An Archimedean copula has the next representation:

$$C(u, v, \delta) = \varphi_\delta^{-1}(\varphi_\delta(u) + \varphi_\delta(v))$$

where φ is a convex, decreasing function defined in (0, 1] with $\varphi(1) = 0$. Some examples of these copulas are:

Gumbel-Barnett family given by Hutchinson and Lai (1990):

$$C(u, v, \delta) = u + v - 1 + (1-u)(1-v)e^{-\delta \ln(1-u)\ln(1-v)}, \quad \delta \in [0,1] \qquad (1)$$

Frank family (1979):

$$C(u, v, \delta) = -\frac{1}{\delta}\ln\left(1 + \frac{(e^{-\delta u} - 1)(e^{-\delta v} - 1)}{e^{-\delta} - 1}\right), \quad \delta \in \Re/\{0\} \qquad (1')$$

Joe's copula (1993):

$$C(u, v, \delta) = 1 - [(1-u)^\delta + (1-v)^\delta - (1-u)^\delta(1-v)^\delta]^{1/\delta}, \quad \delta \ge 1 \qquad (1'')$$

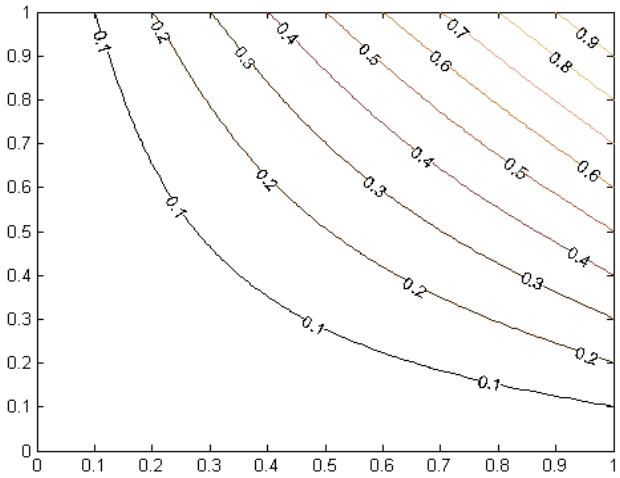


Fig. 1 The level curves of C(1), for $\delta = 0.5$

The bivariate distribution in this case is:

$$F(x,y) = F_1(x) + F_2(x) - 1 + (1 - F_1(x))(1 - F_2(y))e^{-\delta \ln(1-F_1(x))\ln(1-F_2(y))}.$$

If $F_1(\alpha_1, \beta_1)$, $F_2(\alpha_2, \beta_2)$ are Weibull distributed, the bivariate distribution is given by:

$$F(x,y) = 1 - \exp\left(-\left(\frac{x}{\alpha_1}\right)^{\beta_1}\right) - \exp\left(-\left(\frac{y}{\alpha_2}\right)^{\beta_2}\right) + \exp\left(-\left(\frac{x}{\alpha_1}\right)^{\beta_1} - \left(\frac{y}{\alpha_2}\right)^{\beta_2} - \delta\left(\frac{x}{\alpha_1}\right)^{\beta_1}\left(\frac{y}{\alpha_2}\right)^{\beta_2}\right)$$

- Lu and Bhattacharyya (1990) had defined a bivariate Weibull distribution by its survival function :

$$S(x,y) = \exp\left\{-\left[\left(\frac{x}{\alpha_1}\right)^{\beta_1} + \left(\frac{y}{\alpha_2}\right)^{\beta_2} + \delta\left(1 - e^{-\left(\frac{x}{\alpha_1}\right)^{\beta_1}}\right)\left(1 - e^{-\left(\frac{y}{\alpha_2}\right)^{\beta_2}}\right)\right]\right\}, \delta \in [-1,1]$$

with marginals survival functions:

$$S_1(x) = \lim_{y\to 0} S(x,y) = \exp\left(-\left(\frac{x}{\alpha_1}\right)^{\beta_1}\right), x > 0, \quad S_2(y) = \lim_{x\to 0} S(x,y) = \exp\left(-\left(\frac{y}{\alpha_2}\right)^{\beta_2}\right), y > 0.$$

The associated copula for the joint distribution function F(x, y) =S(x, y)-$S_1(x)$-$S_2(y)$ +1 is:

$$C(u,v) = (1-u)(1-v)e^{-\delta uv} - 1 + u + v. \qquad (2)$$

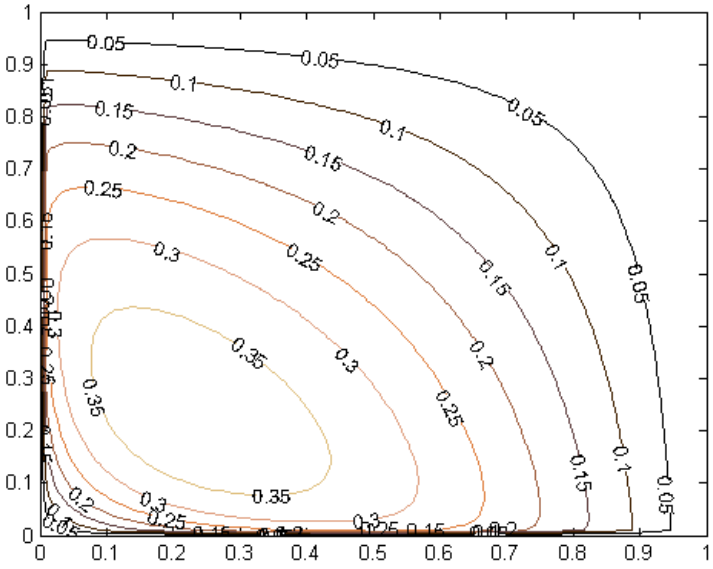


Fig. 2 The level curves of C(1), for $\delta = 0.2$

1. The dependence coefficients

For a copula, the correlation coefficients as Kendall-tau defined by:

$$\tau = 4 \iint_{[0,1]^2} C(u,v) \frac{\partial^2 C}{\partial u \partial v} du dv - 1$$

and Spearman-ro

$$\rho = 12 \iint_{[0,1]^2} C(u,v) du dv - 3$$

are constant.

The linear correlation coefficient r based on the covariance of two variables is not preserved by copulas:

$$r = \frac{M(XY) - M(X)M(Y)}{\sqrt{Var(X)Var(Y)}}$$

where M is the theoretical mean and Var is the theoretical variance.

The tail concentration functions

Right (R) and left (L) tail concentration function can be defined with reference to how much probability is the region near (1, 1) and (0, 0).

These are an intermediate step between correlation coefficients as Kendal, Spearman and copula function itself:

$$L(x) = \frac{C(x,x)}{x}\ ,\ R(x) = \frac{1-2x+C(x,x)}{1-x}$$

$\lambda_u = \lim_{x\to 1} R(x)$ And $\lambda_l = \lim_{x\to 0} L(x)$ are the Tail dependence coefficients. These gave asymptotic measures of the dependence in the tails of bivariate distributions.

For copula (1): $\lambda_u = 0 = \lambda_1 . \lambda_1 = 0$ indicates asymptotic independence in the lower tail. For copula (2): $\lambda_u = 0 = \lambda_1$.

2. The concordance function

Let F be a joint bivariate distribution and G an other joint bivariate distribution.

The concordance function Q is the difference of the probabilities of concordance and discordance between two vectors (X_1,Y_1) and (X_2,Y_2) of continuous random variables with different joint distributions F and G, but with common margins F_1 and F_2.

The function depends on the distributions of (X_1,Y_1) and (X_2,Y_2) only through their copulas.

Let C_1 and C_2 the copulas associated of the vectors (X_1,Y_1) and (X_2,Y_2), so that: $F(x,y)=C_1(F_1(x),F_2(y))$; $G(x,y)=C_2(F_1(x),F_2(y))$. In this case:

$$Q = P[(X1-X2)(Y1-Y2) > 0]- P[(X1-X2)(Y1-Y2) < 0] = 4\iint_{[0,1]^2} C_2 dC_1 - 1 .$$

3. Parametric and semi-parametric estimation procedure

Copula models are used when the association between variables is important. In this case, the effect of the dependence structure is separated from that of the marginals. Two strategies could be envisaged.

Let $(T_{1j}, T_{2j})_{j=\overline{1,n}}$ be a sample of the failure times for the variables T_1 and T_2, where:

$T_1 \sim \text{Weibull}(\alpha_1,\beta_1), T_2 \sim \text{Weibull}(\alpha_2,\beta_2)$ with probability density f_1 and f_2.

The probability density considered for $Weibull(\alpha\ ,\beta)$ is:

$$f(x) = \frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^\beta}\ ,\ x>0.$$

The first method is a two stage estimation method. The 1-st stage involves maximum likelihood for univariate marginals parameters. This procedure is

computationally simpler than estimating all parameters. So, the equations that must be solved are: $\sum_{j=1}^{n}\frac{\partial \ln f_i(t_{ij},\alpha_i,\beta_i)}{\partial\alpha_i}=0$, $\sum_{j=1}^{n}\frac{\partial \ln f_i(t_{ij},\alpha_i,\beta_i)}{\partial\beta_i}=0$ i=1, 2 with solution:

$$\alpha_i=\left(n/\sum_{j=1}^{n}t_j^{\beta_i}\right)^{1/\beta_i} \quad \text{i=1, 2}$$

$$\frac{n}{\beta_i}-n\ln(\alpha_i)+\ln\left(\prod_{j=1}^{n}t_j^{\beta_i}\right)+\left(\frac{1}{\alpha_i}\right)^{\beta_i}\left[\ln(\alpha_i)\sum_{j=1}^{n}t_j^{\beta_i}-\sum_{j=1}^{n}t_j^{\beta_i}\ln(t_j)\right]=0\text{; i=1, 2}$$

and $\sum_{j=1}^{n}\frac{\partial}{\partial\delta}\ln f(t_{1j},t_{2j},\hat\alpha_1,\hat\alpha_2,\hat\beta_1,\hat\beta_2)=0$ where $\frac{\partial^2 F}{\partial t_1\partial t_2}=f$ is the p.d.f.

The second method is a semi-parametric one. The procedure consist of selecting the parameter value that maximize the pseudo-likelihood

$$L(\delta)=\sum_{j=1}^{n}\ln\left[c_\delta(F_{1n}(t_{1j}),F_{2n}(t_{2j}))\right]$$

where F_{in} is the empirical distribution function of i-th variable, i=1,2.

In [5] is proved that the semi-parametric estimator $\hat\delta_n$ is consistent and: $\sqrt{n}\left(\hat\delta_n-\delta\right)$ is asymptotically normal.

The estimation $\hat\delta_n$ is given by the equation:

$$\sum_{j=1}^{n}\frac{\partial}{\partial\delta}\ln\left[c_\delta(F_{1n}(t_{1j}),F_{2n}(t_{2j}))\right]=0 \text{ where } c_\delta(u,v)=\frac{\partial^2 C}{\partial u\partial v}(u,v).$$

In our case for copula (1) this is:

$$\sum_{j=1}^{n}\ln(1-u_j)\ln(1-v_j)=$$

$$\sum_{j=1}^{n}\frac{2\delta\ln(1-u_j)\ln(1-v_j)-1-\ln(1-u_j)-\ln(1-v_j)}{1-\delta\ln(1-u_j)-\delta\ln(1-v_j)-\delta+\delta^2\ln(1-u_j)\ln(1-v_j)}$$

with $u_j=\ln F_{1n}(t_{1j}), v_j=\ln F_{2n}(t_{2j})$.

For the copula (2) $C(u,v,\delta)=(1-u)(1-v)e^{-\delta uv}+u+v-1$ associated to the bivariate Weibull distribution the equation that gives the dependence parameter is:

$$\sum_{j=1}^{n} u_j v_j + \sum_{j=1}^{n} \frac{2u_j + 2v_j - 3u_j v_j - 1 + 2\delta u_j v_j (1-u_j)(1-v_j)}{1+\delta(2u_j + 2v_j - 3u_j v_j - 1) + \delta^2 u_j v_j (1-u_j)(1-v_j)} = 0 .$$

4. Application

We use the recurrent data found at the address **http://www-unix.oit.umass.edu/~statdata/statdata/data/recur.dat**

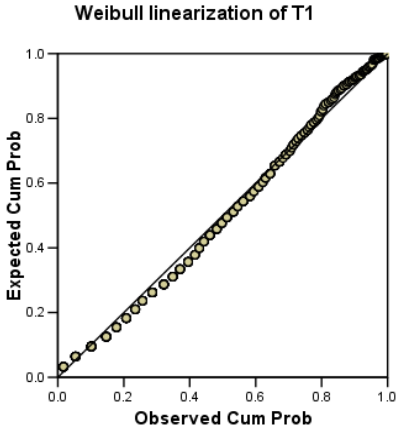
Table 1

ID	AGE	TREAT	T1	T2	CENSOR	EVENT
1	43	0	9	56	1	3
1	43	0	56	88	1	4
1	43	0	0	6	1	1
1	43	0	6	9	1	2
2	43	0	0	42	1	1
2	43	0	87	91	0	3
2	43	0	42	87	1	2
3	41	0	0	15	1	1
3	41	0	15	17	1	2
3	41	0	17	36	1	3
3	41	0	36	112	0	4

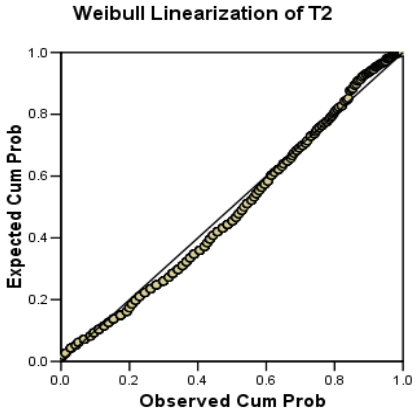
The sample consists in 386 pacients registered with the first time when the disease occurred and the next recurrence of it. These times are T_1 and T_2. From all data, we selected only these that have event = 2.

The concordance with Weibull distribution is established by the linearization method . The values of the parameters are:

$$\alpha_1 = 1.0109524, \beta_1 = 29.222882, \alpha_2 = 1.459987, \beta_2 = 59.11479$$



The figure of linearization are:



The estimated correlation coefficients are:

Table 2

Correlations

		T1	T2
T1	Pearson Correlation	1	.780(**)
	Sig. (2-tailed)	.	.000
	N	386	386
T2	Pearson Correlation	.780(**)	1
	Sig. (2-tailed)	.000	.
	N	386	386

Table 3

Correlations

			T1	T2
Kendall's tau_b	T1	Correlation Coefficient	1.000	.588(**)
		Sig. (2-tailed)	.	.000
		N	386	386
	T2	Correlation Coefficient	.588(**)	1.000
		Sig. (2-tailed)	.000	.
		N	386	386
Spearman's rho	T1	Correlation Coefficient	1.000	.743(**)
		Sig. (2-tailed)	.	.000
		N	386	386
	T2	Correlation Coefficient	.743(**)	1.000
		Sig. (2-tailed)	.000	.
		N	386	386

** Correlation is significant at the 0.01 level (2-tailed).

For copula (1) we define the function:

$$f(\delta)=\sum_{j=1}^{n}\ln(1-u_j)\ln(1-v_j)-\frac{2\delta\ln(1-u_j)\ln(1-v_j)-1-\ln(1-u_j)-\ln(1-v_j)}{1-\delta\ln(1-u_j)-\delta\ln(1-v_j)-\delta+\delta^2\ln(1-u_j)\ln(1-v_j)}$$

and for copula (2) we define the function

$$f(\delta)=\sum_{j=1}^{n}u_jv_j+\sum_{j=1}^{n}\frac{2u_j+2v_j-3u_jv_j-1+2\delta u_jv_j(1-u_j)(1-v_j)}{1+\delta(2u_j+2v_j-3u_jv_j-1)+\delta^2u_jv_j(1-u_j)(1-v_j)}.$$

and we search for the function $f(\delta)$ the root $\delta\in[0,1]$. We present the graphs of $f(\delta)$ in the two cases.

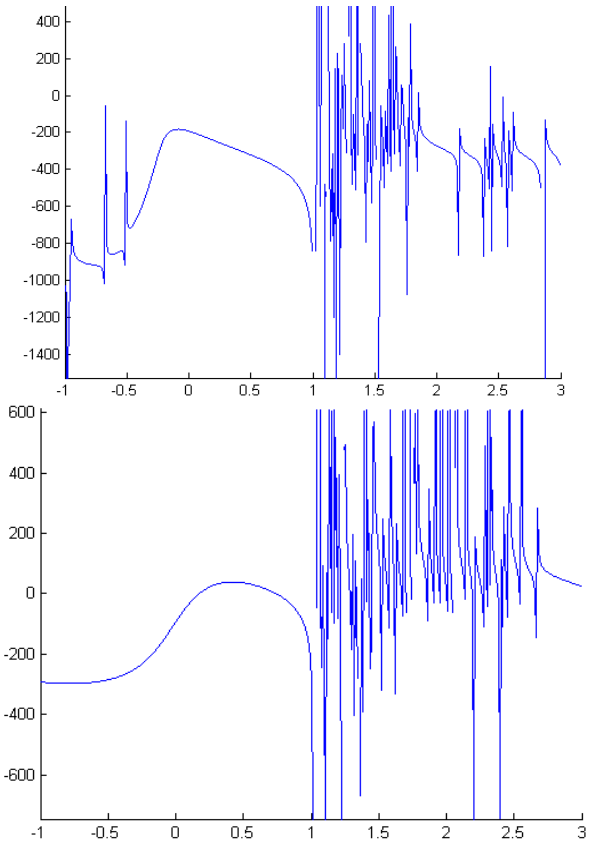


Fig. 3 The function $f(\delta)$ for the two copulas

In the first case one observe that $\delta \notin [0,1]$. This shows to us that for this copula the number of data must be much more. In the second case we find two solutions $\delta_1 \in [0, 0.5]$ and $\delta_2 \in [0.5,1]$. Knowing that the parameter of dependence δ for a such copula is in the interval [-0,20 ; 0,32] we find the value $\delta \cong 0.1706$ with an error or order 10^{-2}.

Conclusions

The semiparametric method for estimating the dependence parameter of a pair of random variables applied for two different bivariate distributions with the same marginals ask different values for the volume of selection. For the first copula that is an Archimedian one the volume of selection must be bigger then for the second copula, that is associated to a bivariate Weibull. This is a reason for which we can not find a value of $\delta \in [0,1]$ for the copula (1).

R E F E R E N C E S

1. *Lawless F.J.* (2003) "Statistical Models and Methods for Lifetime Data", John Wiley & Sons, New York
2. *Genest C., Rivest L.P.* "Statistical Inference Procedures for Bivariate Archimedean Copulas", Journal of American Statistical Association, Sept. 1993, vol.88, no.423, pp.1034-1043
3. *Kalbfleisch J.D, Prentice R.L.* (2002) "The Statistical Analysis of Failure Time Data", John Wiley & Sons, New York
4. *Shih J.H, Louis Th.A.* (1995) "Inference on association Parameter in Copula Models for Bivariate Survival Data", Biometrics 51, pp. 1384-1399
5. *Genest C., Ghoudi K., Rivest L.P.* "A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions", Biometrika 1995, pp.543-552
6. *Nelsen R.B.* "An introduction to Copulas", 1999, Springer Verlag.