A BIVARIATE DISTRIBUTION IN INSURANCE, USING COPULAS

Mariana CRAIU[1], Romeo BERCIA[2]

In aceasta lucrare construim o repartiție bidimensională pentru o problemă de risc în asigurări, utilizând copule Arhimediene. Folosim o metodă semiparametrică pentru determinarea parametrului de dependență între variabilele aleatoare. Cu criteriul Akaike alegem dintre două modele posibile pe cel mai potrivit.

In this paper we establish a bivariate distribution in an insurance risk problem using Archimedean copulas. The dependence parameter in a copula is estimated by a semiparametric method with nonparametric marginal distributions. We choose between two possible models one with Akaike Information Criterion.

Keywords: Archimedean copula, loss adjustement expences, measures of association.

1. Introduction

A loss distribution is the probability distribution of either the loss or the amount paid from a loss event or of the amount paid from a payment event.

Several branches of the management in an insurance company are hit by the same events: a storm can damage houses, highways, cars and injure people; the insurer's risks associated with these are not independent and so the marginal distributions do not describe join risks adequately. With a copula on can separate the joint distribution into two contributions: the marginal distributions of each variable by itself and the copula that combines these into a joint distribution. In the last few years copulas have become very important in insurance and in the management of risks.

But what is a copula? The answer is given by Fisher in 1997 in his article of the update volume of the "Encyclopedia of Statistical Sciences". By him "Copulas are of interest to statisticians for two main reasons: firstly, as a way of studying scale-free measures of dependence and secondly, as a starting point for constructing families of bivariate distributions".

[1] Prof., Dept. of Mathematics III, University "Politehnica" of Bucharest, ROMANIA
[2] Lecturer, Dept. of Mathematics III, University "Politehnica" of Bucharest, ROMANIA

2. Bivariate copulas

Definition [2] A two dimensional copula C is a function with the properties: a) $C:[0,1]\times[0,1]\to\mathbf{R}$; b) $C(u,1)=u,\forall u\in[0,1]$, $C(1,v)=v,\forall v\in[0,1]$; c) $C(u,0)=C(0,v)=0$; d) $C(u_2,v_2)-C(u_2,v_1)-C(u_1,v_2)+C(u_1,v_1)\geq 0$, $\forall u_1,u_2,v_1,v_2\in[0,1]$.

The connection between a copula and a bivariate distribution is given by the following theorem due to Sklar.

Theorem [2] If *C* is a copula, *F* and *G* are univariate distribution functions, then $H(x,y)=C(F(x),G(y))$ is a joint distribution function with margins *F* and *G*. Otherwise if *H* is a joint distribution function with margins *F* and *G* there exists a copula *C* such that $H(x,y)=C(F(x),G(y))$, for all $x,y\in(-\infty,\infty)$. If *F* and *G* are continuous, then *C* is unique.

Using this theorem of Sklar we can define different bivariate distributions generated by copulas for the same marginals.

A special class of copulas are **Archimedean copulas**. These are used in applications due to their simple form and other “nice “ properties. The term of Archimedean copula first appear in two papers by Genest and Mackay (1986).

Archimedean Copulas can be constructed using a generator. A **generator** is a continuous strictly decreasing convex function $\varphi:[0,1]\to[0,\infty)$ with $\varphi(1)=0$. For an Archimedean copula C, $\varphi(C(u,v))=\varphi(u)+\varphi(v)$ and $\varphi(H(x,y))=\varphi(F(x))+\varphi(G(y))$. where *F* and *G* are the marginal distributions of the bivariate distribution *H(x,y)*.

a) Frank’s copulas belong to the Archimedean family and it is defined by:

$$C_\theta(u,v)=-\frac{1}{\theta}ln\left(1+\frac{(exp(-\theta u)-1)(exp(-\theta v)-1)}{e^{-\theta}-1}\right),\quad \theta\neq 0. \tag{1}$$

For them, the generator is the function $\varphi_\theta(t)=-ln\dfrac{exp(-\theta t)-1}{exp(-\theta)-1},\theta\neq 0.$ Some properties of these copulas are:

1. $C_\theta(u,v)=\hat{C}_\theta(u,v),\forall u,v\in[0,1]$, where $\hat{C}_\theta(u,v)$ is the survival copula defined by $\hat{C}_\theta(u,v)=u+v-1+C_\theta(1-u,1-v)$;
2. The level curves are convex;
3. $C_{-\theta}(u,v)=u-C_\theta(u,1-v)=v-C_\theta(1-u,v),\forall u,v\in[0,1]$;
4. $\lim\limits_{\theta\to 0} C_\theta(u,v)=uv$,

b) Other Archimedean copula is defined by Clayton (1978). It is:

$$C_\alpha(u,v)=(u^{-\alpha}+v^{-\alpha}-1)^{-1/\alpha},\quad \alpha>0, \tag{2}$$

The generator of this copula is: $\varphi_\alpha(t)=\frac{1}{\alpha}\left(t^{-\alpha}-1\right)$. Hutchinson and Lai call it the Pareto family of copulas and Genest and Rivet call it the Clayton family.

3. Some dependence measures

The dependence measures between two variables *X* and *Y* are defined by the copulas.

a) Kendall's tau can capture nonlinear dependences that are not possible to measure with linear correlation. The Kendall's tau (τ) measure gives the probability of concordance minus the probability of discordance for a pair of observations from (*X*,*Y*). A pair of random variables are concordant if "large" (small) values of one tend to be associated with "large" (small) values of the other. For a pair of discordant random variables "large" values of one are associated with "small" values of the other. Kendall's tau can capture nonlinear dependences that are not possible to measure with linear correlation.

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

For a Frank's copula we have $\tau_\theta=1-\frac{4}{\theta}\left(1-D_1(\theta)\right)$, where D_1 is the Deby function $D_1(\theta)=\frac{1}{x}\int_0^x\frac{t}{e^t-1}dt$. Using the power series expansion $\frac{t}{e^t-1}\approx 1-\frac{t}{2}+\frac{t^2}{12}-\frac{t^4}{720}+\ldots$ we obtain $\tau_\theta\approx\frac{\theta}{9}-\frac{\theta^3}{900}+\ldots$.

b) Spearman rho-"distance" between the distributions of *X* and *Y* as represented by their copula *C* and independence represented by *uv*. It is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors (X_1,Y_1) and (X_2,Y_2), where (X_1,Y_1) follows a bivariate distribution with marginals *F*,*G* and (X_2,Y_2) is a vector with independent components distributed by *F* and *G*.

$$\rho=12\iint_{I^2}C(u,v)dudv-3,\ \rho_\theta=1-\frac{12}{\theta}\left(D_1(\theta)-D_2(\theta)\right)$$

with $D_2(x)=\frac{2}{x^2}\int_0^x\frac{t^2}{e^t-1}dt$. While $\frac{t^2}{e^t-1}\approx t-\frac{t^2}{2}+\frac{t^3}{12}-\frac{t^5}{720}+\ldots.$, it follows that $D_2(\theta)\approx 1-\frac{\theta}{3}+\frac{\theta^2}{24}-\frac{\theta^4}{2160}$ and $\rho_\theta\approx\frac{\theta}{6}-\frac{\theta^3}{450}$.

c) Gini’s measure of association for *X* and *Y* measures the concordance relationship or the “distance” between *C* and the copulas $M(u,v) = min(u,v)$, $W(u,v) = max(u+v-1,0)$. It is defined by

$$\gamma_{X,Y} = 4\left(\int_0^1 C(u,1-u)du - \int_0^1 (u - C(u,u))du\right)$$

for *X*,*Y* continuous random variables whose copula is *C*. γ depends on the copula *C* only through its principal diagonal and secondary diagonal sections.

d) Medial correlation coefficient λ is proposed by Blomquist (1950)

$$\lambda = P((X-\tilde{x})(Y-\tilde{y}) > 0) - P((X-\tilde{x})(Y-\tilde{y}) < 0),$$

where $\tilde{x}$ and $\tilde{y}$ are the medians of *X* and *Y*. If *F* and *G* are the marginals of the continuous variables *X* and *Y*, then $\lambda = 4C(0.5,0.5) - 1$. So $\lambda_\alpha = 4\left(2^{-\alpha-1} - 1\right)^{-1/\alpha}$ and $\lambda_\theta = \frac{4}{\theta} ln \frac{1+exp(-\theta/2)}{2} + 1$.

4. Losses and ALAE

In causality insurances it is common to record the expenses that are directly related to the payment of the loss (the allocated loss adjustment expenses ALAE). ALAE includes all costs, fees and expenses for attorneys, costs for employing experts, for detective services, for photographs diagrams and other. Allocated loss adjustment expenses (ALAE) is the amount of expenses incurred directly as the result of a loss event.

Table 1

Value of Losses - x_j and ALAE - y_j

j	x_j	y_j	u_j	v_j	j	x_j	y_j	u_j	v_j
1	1500	301	1/24	7/24	13	11750	2530	13/24	15/24
2	2000	3043	2/24	16/24	14	12500	165	14/24	4/24
3	2500	415	4/24	10/24	15	14000	175	15/24	5/24
4	2500	4940	4/24	17/24	16	14750	28217	16/24	23/24
5	4500	395	5/24	8/24	17	15000	2072	17/24	12/24
6	5000	25	6/24	1/24	18	17500	6328	18/24	18/24
7	5750	34474	7/24	24/24	19	19833	212	19/24	6/24
8	7000	50	9/24	3/24	20	30000	2172	20/24	13/24
9	7000	10593	9/24	21/24	21	33033	7845	21/24	20/24
10	7500	50	10/24	3/24	22	44887	2178	22/24	14/24
11	9000	406	11/24	9/24	23	63500	12251	23/24	22/24
12	10000	1174	12/24	11/24	24	210000	7357	24/24	19/24

The losses and the ALAE, denoted by *X* and respectively *Y*, are usually strongly positively correlated. We are interested in constructing a bivariate model.

Some losses with ALAE for which we'll try to construct a bivariate model using copulas are given in [1]. The data are presented in Table 1.

4. Estimation of joint bivariate distribution

We use for this application a procedure developed by Genest [3] that is a two stage semi-parametric one. At first stage the marginal distribution functions are estimated by empirical distribution functions and at the second stage the copula's parameters are estimated by the likelihood method. Models in this case are based on parametric copulas but nonparametric marginal distributions.

For a class of parametric copulas $C_q(u,v)$ with unknown parameter q the estimator $\hat{q}$ for q is the solution of the equation of maximum verosimility. So $\hat{q} = arg\, max\, L_n(q)$, where $L_n(q) = \sum_{j=1}^{n} ln\, c_q(\hat{F}(x_j),\hat{G}(y_j)),\quad c_q = \frac{\partial^2 C_q}{\partial u \partial v},$ and $\hat{F}(x) = \frac{1}{n}\sum_{j=1}^{n} I(x_j \le x),$ respectively $\hat{G}(y) = \frac{1}{n}\sum_{j=1}^{n} I(y_j \le y),$ are the empirical distributions of the marginals. For copulas of Frank and Clayton, the condition $\frac{dL_n}{dq}(q) = 0$ that define the estimators are:

$$\frac{n}{ln\,\beta} - \frac{n\beta}{1-\beta} + \sum_{j=1}^{n}\left(u_j + v_j + 2\frac{u_j\beta^{u_j} + v\beta^{v_j} - \beta - \left(u_j + v_j\right)\beta^{u_j+v_j}}{\beta - 1 + \left(\beta^{u_j} - 1\right)\left(\beta^{v_j} - 1\right)} \right) = 0 \tag{3}$$

where $\beta = e^{-\theta}$, and respectively

$$\frac{n}{\alpha+1} + \sum_{j=1}^{n}\left(\frac{ln\left(u_j^{-\alpha} + v_j^{-\alpha} - 1\right)}{\alpha^2} - ln\, u_j v_j + \left(\frac{1}{\alpha} + 2\right)\frac{u_j^{-\alpha}\, ln\, u_j + v_j^{-\alpha}\, ln\, v_j}{u_j^{-\alpha} + v_j^{-\alpha} - 1} \right) = 0 \tag{4}$$

where $u_j = \hat{F}(x_j), v_j = \hat{G}(y_j), j = \overline{1,n}$, are given in the Table 1.

Solving (3) and (4), we obtain $\hat{\beta} = 0.1388$ and $\hat{\alpha} = 0.4582$. These two copula models are compared using the Akaike Information Criterion $AIC = 2r - 2L_n(q)$ and the Bayes Information Criterion $BIC = r\, ln\, n - 2L_n(q)$, where n is the sample size and r the number of estimated parameters. The Bayes Information Criterion penalizes more strictly for over fitting a model. The best model minimizes *AIC* and *BIC*. While $AIC(\hat{\beta}) = -0.6009, AIC(\hat{\alpha}) = 0.3455$ and $BIC(\hat{\beta}) = 3.7554, BIC(\hat{\alpha}) = 4.7016$, we'll select Frank's model.

Chen [4] propose a pseudo-likelihood ratio test (based on the Akaike Information Criterion) for selecting semi parametric copula models in which the marginal distributions are unspecified and the copula function is parametrized. By this test we find $\sum_{j=1}^{n} ln\frac{c_{\alpha}(u_j,v_j)}{c_{\theta}(u_j,v_j)} = -0.4731 < 0$, which means that Frank's copula is preferred.

5. Conclusions

With Kolmogorov-Smirnov Criterion we could establish that the losses (X) are distributed as Lognormal with $F(x)=\Phi(1.5\,ln\,x+9.29)$ and ALAE variable Y is distributed also as Lognormal with $G(y)=\Phi(2.02\,ln\,y+7.16)$, where $\Phi(u)=\frac{1}{2\pi}\int_{-\infty}^{u} exp(-s^2/2)ds$.

From this results that we can accept that the bivariate distribution of (X,Y) is given by

$$P(X \le x, Y \le y) = H(x,y) = -\frac{1}{\theta} ln\left(1+\frac{(exp(-\theta F(x))-1)(exp(-\theta G(y))-1)}{exp(-\theta)-1}\right),$$

$\theta = 1.9747$

REFERENCES

[1]. *S.A. Klugman*, *H.H. Panjer*, *G. Willmot*, ”Loss Models: From Data to Decisions”, John Wiley,New York, 1998.

[2]. *R.B. Nelsen*, ”An Introduction to Copulas”, Springer, 1999.

[3]. *C.Genest*, *K. Choudi*, *L.P. Rivet*, ”A semiparametric estimation of dependence parameters in multivariate families of distributions”, Biometrika, **82**(3), 1995, pp. 543-552.

[4]. *X. Chen*, *Y. Fan*, ”Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection”, The Canadian Journal of Statistics, **33**(3), 2005, pp. 389-414.

[5]. *I. Purcaru*, *I. Mircea*, *Gh. Lazar*, ”Asigurări de persoane şi de bunuri”, Editura Economică”, Bucureşti, 1998

[6]. *W.-R. Heilmann*, “Fundamentals of Risk Theory”, Verlag Versicherungswirtschaft, Karlsruhe, 1988.

[7]. *M. Evans*, *N. Hastings*, *B. Peacock,* ”Statistical Distributions”, John Wiley & Sons, Inc., New Zork, 1993