

Parametric and non-parametric estimate of bivariate survival functions: the copula approach

Silvia Angela Osmetti

Abstract In this paper we discuss the problem on parametric and non parametric estimation of the distributions generated by the Marshall-Olkin copula. This copula comes from the Marshall-Olkin bivariate exponential distribution used in reliability analysis. Through this copula we can extend the Marshall-Olkin distribution in order to construct several bivariate survival functions. The cumulative distribution functions of these distributions are not absolute continuous functions and they unknown parameters are often not be obtained in explicit form. In particular we consider the IFM method to find the Marshall-Olkin copula estimator, presenting the copula likelihood function. We compare this procedure with a non parametric estimator of the copula, the bivariate empirical copula, used to evaluate the copula goodness of fit. The estimate procedures described are verified through several simulation. One data-set is analyzed for a illustrative purpose.

Key words: Empirical Copula, Marshall-Olkin copula, Reliability analysis

1 The Marshall-Olkin Copula

The Marshall-Olkin bivariate exponential distribution (MOBED) is used in reliability analysis to study complex systems in which the life time random variables of the components are not independent (see Marshall and Olkin 1976). In multivariate distributions the dependence structure existing among the marginal random variables (r.v.s) is described by the means of copula, which is an helpful tool for handling multivariate probability law with given univariate marginals (see Nelsen [8]). The copula approach makes it possible to generalize the MOBED and generate distributions with different marginals but the same dependence structure.

A bivariate copula is a function $C : [0, 1] \times [0, 1] \rightarrow [0, 1]$, such that by an appropri-

Silvia Angela Osmetti, Department of Statistical Science, Università Cattolica del Sacro Cuore, e-mail: silvia.osmetti@unicatt.it

ate extension of its domain on R^2 , has all the properties of a cumulative distribution function (c.d.f.). It can be treated as the bivariate c.d.f. of a r.v. (U, V) with uniform marginals in $[0, 1]$

$$C(u, v) = P(U \leq u, V \leq v) \quad 0 \leq u, v \leq 1.$$

To better understand the copula model we recall the Sklar's Theorem.

Theorem 1 (Sklar). *Let (X, Y) be a bivariate r.v. with c.d.f. $F_{X,Y}(x, y)$ and marginals $F_X(x)$ and $F_Y(y)$; there exist a copula function $C : I^2 \rightarrow I$ such that $\forall x, y \in \mathcal{R}$*

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \quad (1)$$

If $F_X(x)$ and $F_Y(y)$ are continuous then the copula C is unique. Conversely if C is a copula and $F_X(x)$ and $F_Y(y)$ are the marginal distribution functions, then the $F_{X,Y}(x, y)$ function in (1) is a joint distribution function.

This theorem tells us that, given a bivariate c.d.f. and continuous marginals, there is a function C such that the bivariate c.d.f. can be univocally expressed as a function of the marginals. Moreover if $F_X(x)$ and $F_Y(y)$ are continuous the copula can be found by the inverse of equation (1)

$$C(u, v) = F_{X,Y}(F^{-1}(u), F^{-1}(v)), \quad u = F_X(x), v = F_Y(y). \quad (2)$$

In the reliability analysis it is often convenient to express a joint survival function $\bar{F}_{X,Y}(x, y)$ as a copula of its marginal survival functions $\bar{F}_X(x)$ and $\bar{F}_Y(y)$; more specifically exists a function $\hat{C} : [0, 1] \times [0, 1] \rightarrow [0, 1]$, such that

$$\bar{F}_{X,Y}(x, y) = \hat{C}(\bar{F}_X(x), \bar{F}_Y(y)) \quad (3)$$

The function $\hat{C}$ is a copula to which refer as the survival copula.

If we consider the simple case of exchangeable marginal r.v.s, the survival Marshall-Olkin copula (MOC) is

$$C(u, v) = uv \min(u^{-\theta}, v^{-\theta}) \quad (4)$$

where the parameter θ has values in $[0, 1]$ and reflexes the dependence structure between the marginal r.v.s, that are *non-negatively dependent*. If $\theta = 0$ the variables are independent and the copula is called independence copula $C(u, v) = \Pi(u, v) = uv$; if $\theta = 1$ the marginals are co-monotonic and the copula is $C(u, v) = M(u, v) = \min(u, v)$. For different values of θ we find several copulae in the so-called Fréchet-Hoeffding class $\Pi(u, v) \leq C(u, v) \leq M(u, v)$.

By the copula we can calculate the association measures between the variables: the Spearman correlation coefficient $\rho_s = \frac{3}{(4/\theta)-1} \geq 0$ and the Kendall-tau $\tau = \frac{\theta}{2-\theta} \geq 0$. When $\theta \rightarrow 0$ then $\rho_s, \tau \rightarrow 0$; when $\theta \rightarrow 1$ then $\rho_s, \tau \rightarrow 1$. The marginal r.v.s have so positive association. Moreover, the MOC is an extreme value copula because the copula is max-stable (Nelsen [8]): hence $\forall k > 0 \ C(u, v) = C^k(u^{1/k}, v^{1/k}) \ \forall u, v \in I$.

For this reason different marginal distributions in the copula (4) generate new bivariate distributions used in reliability analysis and also in extreme value theory (for example, bivariate Weibull distributions and bivariate extreme value distributions.)

2 The two stage maximum likelihood estimator

In literature the problem of the estimation of multivariate distributions or copula parameters is usually approached by the maximum likelihood method (see, for example, Shih, Luis [10] Joe [5] and Proschan and Sullo [9]). Iterative procedures are also proposed. However, the c.d.f. generated by the MOC are not absolutely continuous, being a mixture of an absolute continuous distribution and a singular one; for this distributions the maximum likelihood method not always gives a solution (Beims et al. 1972 [1]).

The use of copula suggest the *inference functions of margins* or IFM method (see Joe [5]). With this approach we can estimate the parameters of bi/multivariate distribution in two steps: in the first one we estimate the marginal distribution parameters and then in the second step we estimate only the copula parameters. Let $\underline{X} = (X_1, X_2, \dots, X_k)$ be a multivariate random variable with cumulative distribution function

$$F_{\underline{X}}(\underline{x}, \lambda_1, \dots, \lambda_k, \theta) = C(F_{X_1}(x_1, \lambda_1), F_{X_2}(x_2, \lambda_2), \dots, F_{X_k}(x_k, \lambda_k), \theta). \quad (5)$$

where F_{X_j} are continuous marginal c.d.f. with density function $f_{X_j}(x_j, \lambda_j)$ dependent on λ_j parameter in parametric space Λ_j , and C is a copula with parameter $\theta \in \Theta$ and density c . We consider the k log-likelihood functions $l_j(\lambda_j, \underline{x}) = \sum_{i=1}^n \ln f_{X_j}(x_{ij}, \lambda_j)$ with $j=1, 2, \dots, k$ for the univariate r.v.s and the log-likelihood function for the joint distribution

$$l(\underline{\lambda}, \theta, \underline{x}) = \sum_{i=1}^n \ln f(x_{i1}, \dots, x_{ik}, \lambda_1, \dots, \lambda_k, \theta)$$

with $f(\underline{x}, \lambda_1, \dots, \lambda_k, \theta) = c(F_{X_1}(x_1, \lambda_1), \dots, F_{X_k}(x_k, \lambda_k), \theta) \prod_{j=1}^k f_{X_j}(x_j, \lambda_j)$.

The IFM method consists of k separate optimizations of the univariate likelihood to get $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k$, followed by an optimization of the condizionate log-likelihood $l(\theta, \underline{x} | \hat{\underline{\lambda}}) = \sum_{i=1}^n \ln c(F_{X_1}(x_{i1}, \hat{\lambda}_1), \dots, F_{X_k}(x_{ik}, \hat{\lambda}_k), \theta)$, to get $\hat{\theta}$. Under regularity condition, $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k$ and θ are so the solutions of

$$\text{Step I: } \frac{\partial l_j(\lambda_j, \underline{x})}{\partial \lambda_j} = 0 \quad \forall j \quad ; \quad \text{Step II: } \frac{\partial l(\theta, \underline{x} | \hat{\underline{\lambda}})}{\partial \theta} = 0.$$

This procedure is computationally simpler than the usual maximum likelihood one, where all parameters are estimated simultaneously. So, the use of copula can made

a simplification of the estimation method, reducing the computational difficulty and being also less time-consuming. The IFM estimators are usually different from the ML estimators but still have good properties. The properties of consistency and asymptotical efficiency are studied. Their asymptotic distribution is normal with variance-covariance matrix equal to the inverse of Godam Information Matrix (see Joe [5]).

2.1 The likelihood function of the Marshall-Olkin copula

With the IFM method we estimate the parameters of the bivariate survival function generated by the MOC in (4). We present the copula likelihood function and we find the copula estimator with a simpler analytic form.

To construct the log-likelihood function we need the copula density c . The copula in (4) has an absolutely continuous part and a singularity for $u = v$ with positive probability, such that C can be expressed as a liner combination of the absolutely continuous component C_a and the singular one C_s :

$$C(u, v) = \frac{2 - 2\theta}{2 - \theta} C_a(u, v) + \frac{\theta}{2 - \theta} C_s(u, v) \quad (6)$$

with $C_s(u, v) = [\min(u^\theta, v^\theta)]^{\frac{2-\theta}{\theta}}$. Using (6), we can find the density function of the copula, with respect to a dominating measure, that is

$$c_\theta(u, v) = \begin{cases} (1 - \theta)^{\frac{1}{uv}} C_\theta(u, v) & u > v \\ (1 - \theta)^{\frac{1}{uv}} C_\theta(u, v) & u < v \\ \theta^{\frac{1}{u}} C_\theta(u, v) & u = v \end{cases} \quad (7)$$

with $0 \leq u, v \leq 1$ and $0 < \theta < 1$. We draw now a sample of n iid observations from the r.v. (X, Y) generated by the MOC and marginal distributions dependent both on a vector of parameters $\underline{\lambda}$. Using now the IFM, we estimate in the first step the vector $\underline{\lambda}$ and in the second step the copula parameter θ . Therefore, let $\hat{u}_i = F_X(x_i, \hat{\underline{\lambda}})$ and $\hat{v}_i = F_Y(y_i, \hat{\underline{\lambda}})$ we maximize, respect to $\theta \in (0, 1)$, the likelihood function

$$L(\theta | \hat{\underline{u}}, \hat{\underline{v}}) \propto (1 - \theta)^{n_1 + n_2} \theta^{n_3} \prod_{i=1}^n C_\theta(\hat{u}_i, \hat{v}_i).$$

By the logit transformation of θ , $\psi = \ln\left(\frac{\theta}{1-\theta}\right)$, $-\infty < \psi < +\infty$, we solve the maximization problem $\psi \in \Psi$, $\max_{\psi \in \Psi} l(\psi | \hat{\underline{u}}, \hat{\underline{v}})$ with respect to the new parameter, where l is the likelihood function conditioned with respect to $\hat{\underline{u}}$ and $\hat{\underline{v}}$:

$$l(\psi | \hat{\underline{u}}, \hat{\underline{v}}) = k + (n_1 + n_2) \ln(1 - (1 + e^{-\psi})^{-1}) + n_3 \ln((1 + e^{-\psi})^{-1}) + \ln \sum_{i=1}^n C_\psi(\hat{u}_i, \hat{v}_i),$$

where k is a constant. Solving the equation $\vartheta l(\psi|\hat{u}, \hat{v})/\vartheta \psi = 0$ we obtain the ψ estimator with values on the real line, the unique solution of the maximum problem in the parametric space:

$$\hat{\psi} = -\ln \left[\left(n - 2n_3 - S_{\min} + \sqrt{n^2 + S_{\min}^2 - S_{\min}(2n - 4n_3)} \right) / 2n_3 \right]$$

with $n_3 > 0$ and $S_{\min} = \sum_{i=1}^n \min(-\ln(u_i), -\ln(v_i))$. This solution is the unique accepted solution and it is a maximum solution. Thus the maximum likelihood estimator of θ is given by $\hat{\theta} = (1 + e^{-\psi})^{-1}$.

3 Estimation Methods via Empirical Copula

The Empirical Copula was introduced by Deheuvels (1979, 1981) as empirical dependence function. The empirical bivariate function approximates the copula using the order statistics. Let $\{(x_i, y_i) | i = 1, \dots, n\}$ denote an i.i.d. sample from a continuous bivariate r.v.. Let $x_{(i)}$ and $y_{(j)}$, $1 \leq i, j \leq n$ be the order statistics from the sample. An Empirical Copula C_n is the function

$$C_n \left(\frac{i}{n}, \frac{j}{n} \right) = \frac{1}{n} \sum_{k=1}^n I_{\{x_k \leq x_{(i)}, y_k \leq y_{(j)}\}} \quad (8)$$

where I denotes indicator function. The empirical copula frequency is $c_n \left(\frac{i}{n}, \frac{j}{n} \right) = \frac{1}{n}$. Therefore

$$C_n \left(\frac{i}{n}, \frac{j}{n} \right) = \sum_{p=1}^i \sum_{q=1}^j c_n \left(\frac{p}{n}, \frac{q}{n} \right).$$

The C_n can be used to evaluate the goodness of fit of the copula through the distance between a copula and the empirical one. Generally, the smallest distance to the reference copula implies the best fit. The comparison is defined by the Cramér von Mises statistics

$$\int_{[0,1]^2} n[C_n(u, v) - c_{\theta}(u, v)]^2 d_n(u, v) = \sum_{i=1}^n [C_n(u, v) - C_{\theta}(u, v)]^2$$

where C_n denotes the Deheuvels copula and C the compared copula. Genest *et al.* [3] used the parametric bootstrap procedure to obtain approximate p-values. Other goodness-of-fit measures are based on Kolmogorov and Anderson-Darling distances (see Malevergne *et al.* [6]).

4 Simulation and real example

The IFM performance estimation procedure is evaluated by several Monte Carlo simulations, generating 2000 samples with sample size $n = 100, 500, 1000$ and considering the MOBED. This model is generated by the MOC with different parameter values $\theta = 0.1, 0.7, 0.9$ and two exponential marginal r.v.s with equal rate of failure parameters $\lambda_1^* = \lambda_2^* = 0.7, 1.3, 2$. The θ values correspond to the situation of low, mean and high positive association. The goodness of the estimators is evaluated calculating the bias (B) and the mean square error (MSE). In Table 1 the results

Table 1 IFM maximum likelihood estimates.

θ	λ^*	n	λ_1^*		θ		Time
			B	MSE	B	MSE	
0.9	0.7	100	-0.0064	0.0050	0.0018	0.0005	7.9222
		500	-0.0021	0.0010	0.0006	0.0001	10.272
		1000	-0.0010	0.0005	0.0002	0.0001	11.303
	1.3	100	-0.0103	0.0171	0.0015	0.0005	4.2758
		500	-0.0021	0.0035	0.0002	0.0001	5.5639
		1000	-0.0011	0.0017	0.0002	0.0001	6.0389
	2	100	-0.0182	0.0405	0.0012	0.0006	2.7671
		500	-0.0037	0.0082	0.0003	0.0001	3.5909
		1000	-0.0030	0.0040	0.0003	0.0001	3.9701
0.7	0.7	100	-0.0058	0.0051	0.0047	0.0018	8.2415
		500	-0.0024	0.0009	0.0017	0.0004	10.586
		1000	-0.0013	0.0005	0.0013	0.0002	11.619
	1.3	100	-0.0151	0.0176	0.0033	0.0017	4.4203
		500	-0.0018	0.0034	0.0017	0.0004	5.7073
		1000	-0.0023	0.0018	0.0011	0.0002	6.2695
	2	100	-0.0195	0.0408	0.0014	0.0017	2.8797
		500	-0.0033	0.0080	0.0017	0.0004	3.6914
		1000	-0.0008	0.0043	0.0010	0.0002	4.0629
0.1	0.7	100	-0.0084	0.0050	0.0006	0.0016	8.3777
		500	-0.0001	0.0010	0.0003	0.0003	10.703
		1000	-0.0013	0.0005	-0.0001	0.0002	11.665
	1.3	100	-0.0128	0.0168	0.0013	0.0016	4.5478
		500	0.0005	0.0032	0.0003	0.0003	5.7704
		1000	-0.0007	0.0017	0.0003	0.0002	6.3067
	2	100	-0.0231	0.0420	0.0016	0.0015	2.9564
		500	-0.0039	0.0077	0.0001	0.0003	3.7378
		1000	-0.0029	0.0040	0.0000	0.0002	4.0892

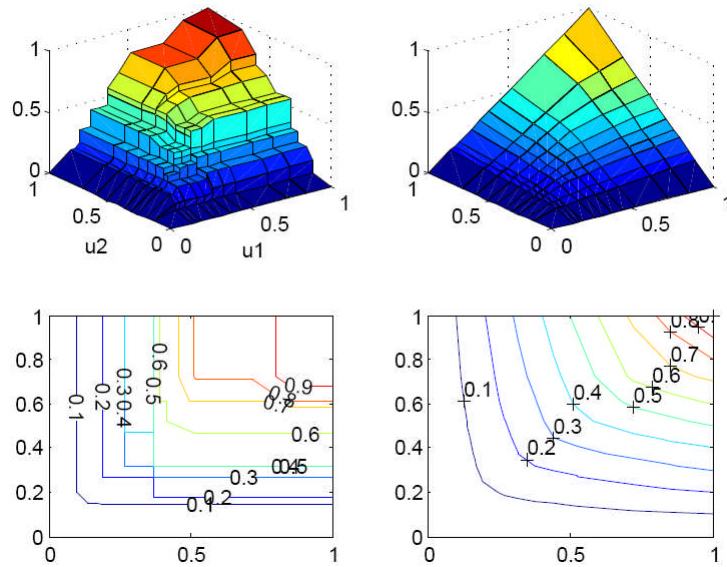
obtained for θ and λ_1^* with the IFM method are shown. We see a good stability of the estimates: the bias and the MSE are usually lower than the one tenth of the real parameter. Once n increase we see an improving of the performances due to a strong decrease of the bias and the MSE.

We conclude analyzing the data set shown in Table 2. It represents the 20 obser-

Table 2 Observed data

Call	X	Y	Call	X	Y
1	8	7	11	1	3
2	10	10	12	11	12
3	10	12	13	12	12
4	5	4	14	6	5
5	6	6	15	8	8
6	1	2	16	11	6
7	2	4	17	1	3
8	4	2	18	2	2
9	7	7	19	6	3
10	6	8	20	5	7

variations related with the time to call 2 elevators in an hotel's reception (the time in minutes between two consecutive pushing button). X represents the time of the calls concerning the first elevator and Y the time for the second elevator. We can have $x_i > y_i$, $x_i < y_i$ and $x_i = y_i$. We suppose that the observations have a MOBED distribution, generated by the MOC with marginal exponential r.v. with rates of failure λ_1^* and λ_2^* . The estimates of the parameters λ_1^* , λ_2^* and θ obtained with the IFM procedure are respectively $\hat{\lambda}_1^* = \hat{\lambda}_2^* = 0.16$ and $\hat{\theta} = 0.52$. The value of θ describes a mean grade of positive association between the marginal r.v.s.

**Fig. 1** Empirical copula and IFM copula estimate

A Kolmogorov-Smirnov test for X and Y doesn't reject the null hypothesis that the data come from a univariate exponential distribution with failure rate equal to 0.16. Moreover we estimate the copula with the empirical copula (8).

The Figure 1 make it possible to evaluate the goodness of fit comparing the IFM copula estimate with the empirical one.

5 Conclusion

In this work we have investigated inferential aspects of copula models. In particular we have considered the IFM criteria to estimate the parameters of a bivariate survival distribution generated by the MOC. We found the estimator of copula parameter and we presented the log-likelihood function. We analyzed the asymptotic properties of the estimators by an experiment simulation. We obtained good results: values of the bias and the mean square error are lower than one then of the real value of the parameter. Moreover, we presented the non parametric copula estimator via empirical copula. A Matlab program was developed to construct the empirical copula. Finally one data set was analyzed.

References

1. Beims B., Bain L.J., Higgins J.J.: Estimation and hypothesis testing for the parameters of a bivariate exponential distribution, *Journal of American Statistical Association* 67, 927-929 (1972)
2. Deheuvels P.: La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance, *Acad Roy Belg Bul Cl Sci* (5), 65, 274-292, (1979)
3. Genest, C., Remillard, B., Beaudoin, D.: Goodness-of-fit for copulas: A review and a power study. *Insur. Math. Econ.*, 44, 199-213, (2009)
4. V.P. Godame: An Optimum Property of Regular Maximum Likelihood Estimation, *Annals of Mathematical Statistics* 31, 1208-1212, (1960)
5. Joe H.: *Multivariate Models and Dependence Concepts*, Chapman&Hall, Boca Raton, (1997)
6. Malevergne, Y., Soornette, D.: Testing the Gaussian copula Hypothesis for financial asset dependences, *Quantitative Finance* 3, 231-250, (2003)
7. Marshall A.W., Olkin I.: A Multivariate Exponential Distribution. *Journal of the American Statistical Association* 62(317), 30-40 (1967)
8. Nelsen R.B.: *An Introduction to Copulas*, Springer, New York, (2006)
9. Proschan F., Sullo P., Estimating the Parameters of a Multivariate Exponential Distribution, *Journal of the American Statistical Association*, 7, 354, 465-472, (1960).
10. Shih J.H., Louis T.A.: Inferences on the Association Parameter in Copula Models for Bivariate Survival Data, *Biometrics*, 51, 4, 1384-1399, (1995)