INTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTERS IN SIMULATION
DOI: 10.46300/9102.2022.16.10 Volume 16, 2022

Archimedean copulas and goodness of fit test

Kazemi Rad, A.M., Golshani, L., Najjari, V.*,
Department of statistics, Faculty of Science, Islamic Azad University, Central Tehran Branch, Iran
Department of mathematics, Islamic Azad university, Maragheh branch, Maragheh, Iran
*fnajjary@yahoo.com

Abstract- The main endeavor of this paper is to compare the result of parameter estimation for Archimedean copulas by using Kendal coefficient and Goodness of fit test. It is seen that on modeling dependency structure of data by the GOF method, at the same time, we are able to estimate parameters and also test the compatibility of copulas to data.

Keywords- Copulas, Archimedean copulas (Ac), Kendal coefficient, Goodness of fit test (GOF test).

I. Introduction

Copulas are used in modeling the dependence structure between variables, this modeling is irrespective of their marginal distributions. On the other hand copulas allow to choose different margins and merge the margins into a genuine multivariate distribution. Sklar (1959) fro the first time used the concept of copula and it has been introduced by in the following way,

A copula is a function $C : [0,1]^2 \to [0,1]$ which satisfies:
(I) for every u, v in [0, 1], $C(u,0) = 0 = C(0,v)$, and $C(u,1) = u$ and $C(1,v) = v$;
(II) for every u_1, u_2, v_1, v_2 in [0, 1] such that $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$.

Sklar's Theorem describes the importance of copulas as follows:

Let X and Y be random variables with joint distribution function H and marginal distribution functions F and G, respectively. Then there exists a copula C such that, $H(x,y) = C(F(x), G(y))$, for all x, y in $\mathbb{R}$. If F and G are continuous, then C is unique. Otherwise, the copula C is uniquely determined on $Ran(F) \times Ran(G)$. Conversely, if C is a copula and F and G are distribution functions, then the function H is a joint distribution function with margins F and G. This means that, one-dimensional margins of joint distribution functions are linked by copulas. For a more formal definition of copulas, the reader is referred to Nelsen (2006).

There are several parametric and non-parametric methods for estimating parameters of copulas. In this paper we compare the result of parameter estimation for Archimedean copulas by using Kendal confident and also Goodness of fit test. It can be an advantage for non-parametric GOF method, as it is able at the same time, to estimate parameters and also test the compatibility of copulas to data.

This paper is constructed as follows: Section 2 discusses related with estimation of copula parameters and copula selection methods. Section 3 explains the GOF method. An application study is given in Section 4 and finally Section 5 summarizes the conclusion of our work.

II. Copula selection methods

In multivariate statistical analysis of copulas one of the main topics is related with statistical inference on the dependence parameter. In the literature several methods proposed for estimation copula parameters. Genest and Rivest (1993) proposing a method which is based on concordance. Genest et al. (1995) proposed fully maximum likelihood (ML), pseudo maximum likelihood (PML). In 2005, Joe discussed on inference function for margins (IFM) and Tsukahara in 2005 proposed minimum distance (MD) method. There are some discussions about these methods by Kim et al. (2007) and also Najjari (2016).

As a result, PML estimator is better than ML and IFM in the most practical situations. simulation study by Kim et al. (2007) carried out that the PML method is conceptually almost the same as the IFM one. By using the PML method, any important statistical insights that would be gained by applying the IFM, would not be loosed. Therefore, the PML estimator is better than those of the ML and IFM in most practical situations. However, in high dimensional copulas ($n > 3$) ML, PML, IFM and MD methods, in time-consuming point of view, require so much computations. In these methods copula density function are used so that increases complexity of calculations, specially for $d > 2$[1] (Yan, 2007). Semi-parametric estimation of copula models based on the method of moments proposed by Brahimi and Necir in 2012. This method is quick and simple, nonetheless, Brahimi and Necir's method has its own complexity.

[1]Dimensions

After estimating copula parameters, another step is in selecting the right copula that has the best fits to data. In the literature several methods proposed for selecting the best copula some of which are summarized as follows:

Most of methods for selecting right copula are based on a likelihood approach.For example, the Akaike Information Criteria (AIC), Pseudo-likelihood ratio test proposed by Chen and Fan (2005).

There are other methods of selecting the best copula which defines indicators of performance. Genest and Rivest (1993) proposed a method in Archimedean copulas as below,

$$K_\theta(t) = P(C(u, v\ |\theta) < t)$$

with its non-parametric estimation K_n, given by

$$K_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(e_{jn} \le t)$$

where $e_{jn} = (1/n)\sum_{k=1}^{n} \mathbf{1}(X_{1k} \le X_{1j}, ..., X_{pk} \le X_{pj})$. A copula that the function K_θ is closest to K_n, is the best one.

Choosing the best copula with minimizing the distance (L^2-norm, Kolmogorov, etc) from K_θ to the non-parametric estimation K_n suggested by Durrleman et al. (2000). Genest et al. (1995) proposed a GOF test statistic with a non-truncated version of Kendall process,

$$\mathbb{K}_n(t) = \sqrt{n}\{K_n(t) - K_{\theta_n}(t)\}$$

where θ_n denotes a robust estimation of θ. The expression for the statistic is simple and the test has nice properties.

In the p-dimensional copulas, Pollard (1979) presented χ^2 tests as bellow:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{[o_{ij} - e_{ij}]^2}{e_{ij}}$$

where n is the sample size, I, J are the numbers of classes. o_{ij} is the observed frequency of data in class ij and e_{ij} is the theoretical frequency of data for every class ij. Growing number of the classes increases power of the test. Kendall and Stuart(1983) discuss related with arbitrary choice of the subsets that divide the p-dimensional space $[0, 1]^p$.

As examples, Dobric and Schmidt (2004) used this method in a financial application. Çelebioğlu (2003) in modeling student grades relies on this method. Najjari & Ünsal (2012) used this method in modeling meteorological data, Najjari et al. (2014) applied this method in modeling data of the Danube river, also see Şahin Tekin et al. (2014), Kazemi Rad et al. (2021), Najjari et al. (2016) and Bacigál et al. (2015).

III. Goodness of fit test in selection of the right copula

All of the mentioned criteria the Section II. rely on previous estimation of an optimal parameter set of copulas to select the right copula. In this section, a method is described in selecting the right copula which is independent of the chosen optimal parameter. On the other hand, at the same time, it is able to estimate the copula parameters and also it is able to select the right copula. For two dimensional data (X, Y) thid method is described as follows:

Time dependency of data are tested, then random samples (X_1, Y_1), $\cdots$, (X_n, Y_n) are converted into normalized ranks in the usual fashion by setting $U_l = rank(X_l)/n$ and $V_l = rank(Y_l)/n$ for each $l \in \{1, 2, \cdots, n\}$. Then the data are grouped into $r \times r$ contingency table by using $r = round(\sqrt[4]{n})$.

Let observed frequencies is matrix $O_{I\times J}$, and matrix $E_{I\times J}$ consists of estimation of the expected frequencies $(I = J = r)$ and in the matrix O, o_{ij} is an element of ith row and jth column, and e_{ij} is the expected value of the o_{ij}, which is calculated by multiplying the number of observations n with appropriate theoretical frequency estimated with copulas, where $i, j \in \{1, 2, \cdots, r\}$. Let copula parameters θ be in the form $\theta = (\theta_1, \theta_2, ..., \theta_s)$, where s is the number of copula parameters. C_θ for some $\theta \in \Theta$ (where Θ is parameters space), expected frequencies $e_{jk}(\theta)$ are computed for the contingency table. So $e_{ij}(\theta)$ is a function of the parameters θ and can be calculated as follows

$$e_{ij}(\theta) = n\times[C_\theta(u_i, v_j) - C_\theta(u_{i-1}, v_j) - C_\theta(u_i, v_{j-1}) + C_\theta(u_{i-1}, v_{j-1})]$$

where $i, j \in \{1, 2, \cdots, r\}$ and $e_{11}(\theta) = n \times C_\theta(u_1, v_1)$. So the standard GOF statistic value is as follows,

$$h(\theta) = \chi^2_\theta = \sum \frac{(o_{ij} - e_{ij}(\theta))^2}{e_{ij}(\theta)}. \tag{1}$$

In order to determine the right copula that fits the best to data, obviously expected frequencies estimated by copula must be close to observed frequencies. Clearly this fact occurs in point $\hat{\theta}$ (estimation of θ) that minimizes χ^2_θ in (1).

Meanwhile, if $\chi^2_{\hat{\theta}}$ in (1) yields a low p-value by reference to the chi-squared distribution with $(r-1)\times(r-1)$ degrees of freedom, then $H_0 : C_{\hat{\theta}} \in C_\theta$ is rejected. On the other hand in the range of copula parameters minimum point of the function $h(\theta)$, is an estimating of the copula parameters and also a way in choosing the right copula that had best fits to data. Without loss of generality it can be assumed that the minimum value is accrued only on a single point of its parameters range and calculating minimum value of the function $h(\theta)$ of course is not complicated and easily applicable in multiparameter copulas. See also Najjari (2016), Kınacı et al.(2016), .

IV. Application

In our application we use minimum and maximum QFE data from 2010-2020 in Tehran-Iran which are 420 data and are available online at I.R OF IRAN METEOROLOGICAL ORGANIZATION (IRIMO). We remember that, QFE is the barometric altimeter setting that will cause an altimeter to read zero when at the reference datum of a particular airfield (in general, a runway threshold). In ISA temperature conditions the altimeter will read height above the datum in the vicinity of the airfield.

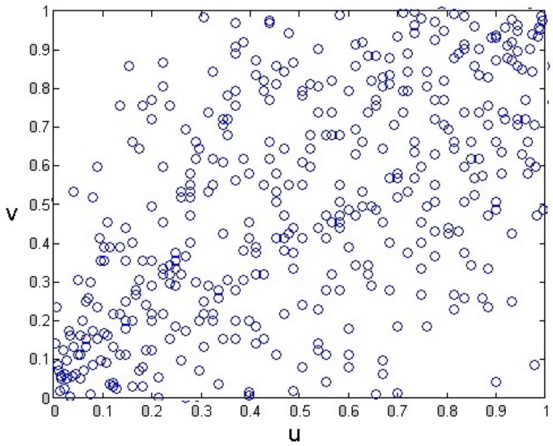

Fig. 1: Scatterplots for minimum and maximum QFE data.

We consider six most widely used Archimedean families of copulas (Table 1, 2): Clayton, Gumbel and A12, A14, A15, A18 (this copula families numbered as 4.2.12, 4.2.14, 4.2.15, 4.2.18 respectively in the Nelsen's book [16]). In continue, related QFE data is arranged and has been divided to total data sample size plus one, as below (see [6], [2]):

$$u_i = \frac{R(x_i)}{n+1} \quad , \quad v_i = \frac{R(y_i)}{n+1} \quad , \quad i = 1, 2, 3, ..., n$$

where x_i is the minimum, and y_i is the maximum QFE data and $R(x_i)$ and $R(y_i)$ are (in ascending order) the rank of related data. Our final data (u and v) will be in interval $(0, 1)$. Figure 1 shows the scatterplots of the final data. Kendall's tau for this data is $\tau = 0.3894$. To apply GOF test, with using relation $\sqrt[4]{n}$, (n number of data), (see [2], [15]), we divided the range of two variables uniform transformation into 4 intervals each, therefore $df = 9$, and the critical point in GOF test is $\chi^2_{0.05,df} = 16.9190$. As Kendall's tau for this data is $\tau = 0.3894$, for each copulas family nonparametric estimation of family parameter applied, and then calculated the value of χ^2 test statistic. Result are in Table 3. In the GOF test method, GOF test statistic, we posed as a function of copula parameter, and then we calculated the minimum value of this function (see Figure 2). Hereby we will estimate copulas parameter and also we will find the right copula between copula families (see Table 3).

Table 1: Definition and parameter domain of the copulas used in this paper.

Family	$C(u,v;\theta)$	θ interval
Clayton	$\max([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0)$	$[-1,\infty) - \{0\}$
Gumbel	$\exp(-[(-lnu)^{\theta} + (-lnv)^{\theta}]^{1/\theta})$	$[1,\infty)$
A12	$(1 + [(u^{-1} - 1)^{\theta} + (v^{-1} - 1)^{\theta}]^{1/\theta})^{-1}$	$[1,\infty)$
A14	$(1 + [(u^{-1/\theta} - 1)^{\theta} + (v^{-1/\theta} - 1)^{\theta}]^{1/\theta})^{-\theta}$	$[1,\infty)$
A15	$max\ ((1 - [(1 - u^{1/\theta})^{\theta} + (1 - v^{1/\theta})^{\theta}]^{1/\theta})^{\theta}, 0)$	$[1,\infty)$
A18	$max\ (1 + \theta/ln(e^{\theta/(u-1)} + e^{\theta/(v-1)}), 0)$	$[2,\infty)$

Table 2: Kandell's tau and it's domain of the copulas used in this paper.

Family	Kendall's tau	Kendall's tau interval
Clayton	$\frac{\theta}{\theta+2}$	$(0,1]$
Gumbel	$\frac{\theta-1}{\theta}$	$[0,1]$
A12	$1-\frac{2}{3\theta}$	$[1/3,1]$
A14	$1-\frac{4}{2+4\theta}$	$[1/3,1]$
A15	$1+\frac{4}{2-4\theta}$	$[-1,1]$
A18	$1-\frac{4}{3\theta}$	$[1/3,1]$

Table 3: Estimated copula parameter and GOF test statistic value.

Family	GOF		Nonparametric method	
	estimated parameter	χ^2	estimated parameter	χ^2
Clayton	**0.91**	**46.287**	**1.28**	**53.4373**
Gumbel	**1.59**	**8.4987**	**1.64**	**8.8204**
A12	**1.11**	**40.9745**	**1.09**	**41.0929**
A14	**1.23**	**34.1538**	**1.14**	**35.8295**
A15	**2.08**	**13.7028**	**2.14**	**14.3921**
A18	**2.21**	**1488.1**	**2.18**	**1507.6**

With respect to Table 3, obviously nonparametric estimation and the new method, selected Gumbel and A15 family as right copulas, also Gumbel copula fits better to data, because it has lowest test statistic value. In Gumbel family, nonparametric estimation of copula parameter is $\theta = 1.64$ and χ^2 test statistic value is 8.8204, while the new method estimated copula parameter as $\theta = 1.59$ and χ^2 test statistic value as 8.4987. In A15 family, nonparametric estimation of copula parameter is $\theta = 2.14$ and χ^2 test statistic value is 14.3921, while the new method estimated copula parameter as $\theta = 2.08$ and χ^2 test statistic value as 13.7028.

V. Conclusion

In this study we used GOF test statistic, as a function on copula parameter. By this technique it is possible to have an estimation for copulas parameters. This means we don't need to rely on previous estimation of an optimal parameter set. Then with calculating minimum point of this function in the range of copula parameter, we will reach to both aims, estimating the copula parameter and choosing the right copula that fits the best to data.

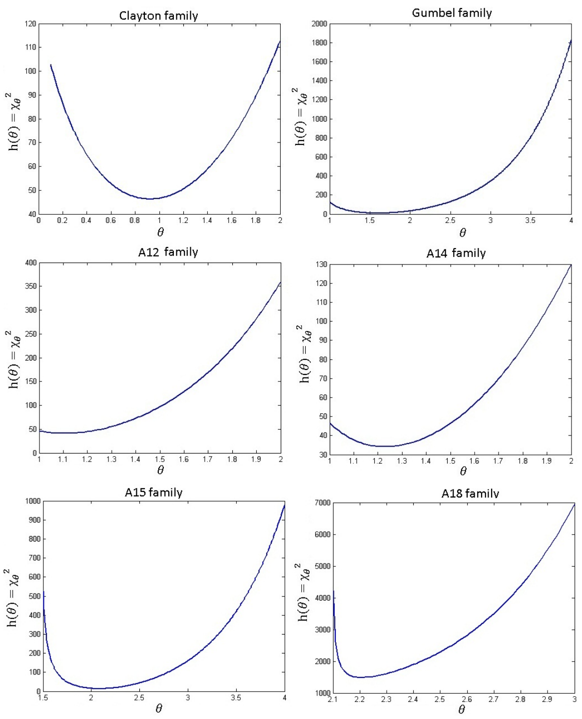

Fig. 2: GOF test statistic, as a function of copula parameter.

References

[1] X. Chen., Y. Fan., Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection, *La Revue Canadienne de Statistique.* 33 (3)(2005) 389–414.

[2] S. Çelebioğlu., Archimedean copulas And An Application, *Selcuk University Journal of Science*, 22 (2003) 43–52.

[3] J. Dobric., F. Schmidt., Testing goodness of fit for parametric families of copulas application to financial data. Seminar of economic and social statistics, *University of Cologne.* 2004.

[4] V. Durrleman., A. Nikeghbali., T. Roncalli., Which copula is the right one? Working document, Groupe de Recherche Opérationnelle, *Crédit Lyonnais*, 2000.

[5] C. Genest., K. Ghoudi., L.P. Rivest., A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, 82 (1995) 543–552.

[6] C. Genest., L. P. Rivest., Statistical Inference Procedures for Bivariate Archimedean copulas, *Journal of the American Statistical Association*, 88:423 (1993) 1034–1043.

[7] H. Joe., Asymptotic efficiency of the two-stage estimation method for copula-based models, *J. Multivariate Anal*, 94 (2005) 401–419.

[8] A.M. Kazemi Rad, L. Golshani, V. Najjari, Archimedean copulas and several methods on constructing generators, *WSEAS TRANSACTIONS on SYSTEMS*, 20 (2021) 228–231.

[9] M. Kendall., A. Stuart., The Advanced Theory of Statistics, *fourth ed. Oxford University Press, NewYork*, vol 2, 1983.

[10] G. Kim., M.J. Silvapulle., P. Silvapulle., Comparison of semiparametric and parametric methods for estimating copulas, *Comm. Statist. Simulation Comput*, 51 (2007) 2836–2850.

[11] H. Kınacı, V. Najjari, I. Alp., Using data envelopment analysis and stochastic frontier analysis methods to evaluate efficiency of hydroelectricity centers, *Gazi University Journal of Science*, 29(1) (2016) 167–176.

[12] V. Najjari., The Dependence Structure Between the Monthly Minimum and Maximum Barometric Data in Iran, *Gazi University Journal of Science*, 29(3):(2016) 593–597.

[13] V. Najjari, H. Bal, F. Özturk, I. Alp., Stochastic frontier models by copulas and an application, *Scientific Bulletin*, 78(1):(2016) 31–41.

[14] V. Najjari., T. Bacigál., H. Bal., An Archimedean Copula Family with Hyperbolic Cotangent Generator, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(5):(2014) 761–768.

[15] V. Najjari., M. G. Ünsal., An Application of Archimedean Copulas for Meteorological Data, *G. U. Journal of Science*, 25:2 (2012) 301–306.

[16] R. B. Nelsen., An introduction to copulas, *Springer, New York*, Second edition, 2006.

[17] D. Pollard., General chi-square goodness-of-fit tests with data-dependent cells. *Z. Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50 (1979) 317–331.

[18] A. Sklar., Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, 8 (1959) 229–231.

[19] S. T. Şahin Tekin., V. Najjari., H. H. Örkcü., Simulation study on copulas, *Sahand Communications in Mathematical Analysis*, 1(2) (2014) 55–63.

[20] T. Bacigál, V. Najjari, R. Mesiar, H. Bal., Additive generators of copulas, *Fuzzy Sets and Systems*, 264 (2015) 42–50.

[21] H. Tsukahara., Semiparametric estimation in copula models, *Canad. J. Statist*, 33 (2005) 357–375.

[22] J. Yan., Enjoy the joy of copulas: with a package copula, *J. Stat. Software*, 21 (4) (2007) 1–21.