In the Name of God

Proceeding of

the 5th Seminar on

Copula Theory and its Applications

Department of Statistics

and

Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad,
Mashhad, Iran

30-31 Jan, 2019.

Preface

The series of biennial workshops on copula theory which took place in Ferdowsi University of Mashhad (2011 and 2013), Shahid Bahonar University of Kerman (2015) and Yazd University (2017) with an emphasis on application in engineering sciences, agricultural sciences, actuarial science, finance, reliability, survival analysis, economics and etc. is the result for the decision of the scientific committee of the Ordered and Spatial Data Center of Excellence (OSDCE) of Ferdowsi University of Mashhad (FUM) to organize workshops and seminars every two years. This seminar is sponsored by the department of statistics, OSDCE of FUM, Islamic world Science Citation database (ISC), Iranian Statistical Society and Actuarial Society of Iran to provide suitable facilities for academics to have efficient research cooperation and will be held at Faculty of Mathematical Sciences of FUM at 30 and 31 Jan. 2019. We hope all of the seminar committees provide a suitable satisfactory atmosphere for the participants. After the first call of the seminar, 20 papers were accepted as oral presentations and 7 as poster presentations by the referees and scientific committee. The attendants and participants in the seminar are in summary 40 people which are professors, students and researchers of different institutes around Iran. Finally, we would like to extend our sincere gratitude to the Research Council of FUM, the administration of Faculty of Mathematical Sciences, the OSDCE, the Islamic world Science Citation center, the Iranian Statistical Society, Actuarial Society of Iran, the scientific committee, the organizing committee, the referees, and the students and staff of the department of statistics of FUM for their kind cooperation.

Mohammad Amini (Chair)
Jan, 2019

Topics

The aim of the seminar is to provide a forum for presentation and discussion of scientific works covering theories and methods such as:

- Methods of copula construction
- Copula functions and dependence concepts
- Dependence modelling using copula function
- Inference based on copula
- Application of copula in spatial, survival, reliability, engineering, hydrological, meteorological, agricultural, finance, economic data and etc.

Scientific Committee (Alphabetical order)

1. Alamatsaz, M.H., University of Isfahan
2. Amini, M., Ferdowsi University of Mashhad (Chair)
3. Dolati, A., Yazd University
4. Hasanzadeh, A., Shahid Beheshti University
5. Jabbari Nooghabi, H., Ferdowsi University of Mashhad
6. Kheiri, S., Shahrekord University of Medical Sciences
7. Mirhosseini, S.M., Yazd University
8. Mohammadzadeh, M., Tarbiat Modares University
9. Mohtashami Borzadaran, G.R., Ferdowsi University of Mashhad
10. Parham, G.A., Shahid Chamran University of Ahvaz
11. Shams, S., Alzahra University
12. Shishehbor, Z., Shiraz University
13. Zounemat-Kermani, M., Shahid Bahonar University of Kerman

Organizing Committee (Alphabetical order)

1. Ahmadi, J., Ferdowsi University of Mashhad
2. Amini, M., Ferdowsi University of Mashhad (Chair)
3. Dolati, A., Yazd University
4. Doostparast, M., Ferdowsi University of Mashhad
5. Fakoor, V., Ferdowsi University of Mashhad
6. Jabbari Nooghabi, H., Ferdowsi University of Mashhad
7. Jabbari Nooghabi, M., Ferdowsi University of Mashhad
8. Mohtashami Borzadaran, G.R., Ferdowsi University of Mashhad
9. Sadeghpour Gildeh, B., Ferdowsi University of Mashhad

Student Organizing Committee (Alphabetical order)

1. Ahmadi Nadi, A., PhD student in Statistics, Ferdowsi University of Mashhad

2. Amini, E., Undergraduate student in Computer Engineering, Ferdowsi University of Mashhad
3. Dodman, N., PhD student in Statistics, Ferdowsi University of Mashhad
4. Esfahani, M., PhD student in Statistics, Ferdowsi University of Mashhad
5. Hoti, N., PhD student in Statistics, Ferdowsi University of Mashhad
6. Kazempoor, J., PhD student in Statistics, Ferdowsi University of Mashhad
7. Mohammadi, M., PhD student in Statistics, Ferdowsi University of Mashhad
8. Mohtashami Borzadaran, H.A., PhD student in Statistics, Ferdowsi University of Mashhad (Head of student organizing committee)

Contents

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

On Stochastic Comparisons of Extreme Order Statistics from the Proportional Odds Model of Distributions

Bashkar, E. [1] Kundu, A. [2]

[1] Department of Statistics, Velayat University, Iranshahr, Iran
[2] Department of Mathematics, Santipur College, Santipur Nadia, West Bengal, India

Abstract

In this paper, we study stochastic comparison of the smallest and largest order statistics of two heterogeneous random vectors with dependent components having proportional odds marginals and Archimedean copula structure in terms of the usual stochastic order.

Keywords: Archimedean copula, Proportional odds model, Majorization, Usual stochastic order.

1 Introduction

There is an extensive literature on different stochastic orderings among order statistics where the observations come from different family of distributions. Some of these contributions

[1]esmaielbashkar@gmail.com
[2]bapai_k@yahoo.com

are due to [20] , [8], [21], [5], [6], [7] [24], [14], [13], [11], [16] and [9], [12], [2], [23]. A recent review on the topic can be also found in [1]. Suppose order statistics arising from random variables $X_1, \ldots, X_n$ are denoted by $X_{1:n} \leq \ldots \leq X_{n:n}$. Then it is well-known that the kth order statistic of a sample of size n characterizes the lifetime of a $(n-k+1)$-out-of-n system. Thus, the study of lifetimes of k-out-of-n systems is equivalent to the study of the stochastic properties of order statistics. In particular, a 1-out-of-n system corresponds to a parallel system and an n-out-of-n system corresponds to a series system. Recently, some efforts are made to investigate stochastic comparisons on order statistics of r.v.s with Archimedean copulas. See, for example, [3], [16], [15] and [10]. The proportional odds (PO) model introduced by Bennet [4] is a very important model in survival analysis context. In this paper, we study the smallest and largest order statistics from two dependent samples with proportional odds (PO) samples. Throughout this paper, we use the notation $\mathbb{R} = (-\infty, +\infty)$.

Let X and Y be two univariate random variables withsurvival functions $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$, respectively. Random variable X is said to be smaller than Y in the usual stochastic order, denoted by $X \leq_{st} Y$, if $\bar{F}(x) \leq \bar{G}(x)$ for x. For a comprehensive discussion on various stochastic orders, one can see [22]. A real function ϕ is n-monotone on $(a, b) \subseteq \mathbb{R}$ if $(-1)^{n-2}\phi^{(n-2)}$ is decreasing and convex in (a, b) and $(-1)^k \phi^{(k)}(x) \geq 0$ for all $x \in (a, b), k = 0, 1, \ldots, n-2$, in which $\phi^{(i)}(.)$ is the ith derivative of $\phi(.)$. For a n-monotone $(n \geq 2)$ function $\phi : [0, +\infty) \longrightarrow [0, 1]$ with $\phi(0) = 1$ and $\lim_{x \to +\infty} \phi(x) = 0$, let $\psi = \phi,^{-1}$ be the right continuous inverse of ψ, then

$$C_\phi(u_1, \ldots, u_n) = \phi(\psi(u_1) + \ldots + \psi(u_n)), \quad \text{for all } u_i \in [0, 1], i = 1, \ldots, n,$$

is called an Archimedean copula with generator ϕ. Archimedean copulas cover a wide range of dependence structures including the independence copula with generator $\phi(t) = e^{-t}$. For more on Archimedean copulas, readers may refer to [19] and [18].

It is well known that the notion of majorization is extremely useful and powerful in establishing various inequalities. For preliminary notations and terminologies on majorization theory, we refer the reader to [17]. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be two real vectors anf $x_{(1)} \leq \ldots \leq x_{(n)}$ be the increasing arrangement of the components of the vector $\mathbf{x}$.

Definition 1. The vector $\boldsymbol{x}$ is said to be

(i) weakly supermajorized by the vector $\boldsymbol{y}$ (denoted by $\boldsymbol{x} \overset{\text{w}}{\preceq} \boldsymbol{y}$) if $\sum_{i=1}^{j} x_{(i)} \geq \sum_{i=1}^{j} y_{(i)}$ for all $j = 1, \ldots, n$,

(ii) majorized by the vector $\boldsymbol{y}$ (denoted by $\boldsymbol{x} \overset{\text{m}}{\preceq} \boldsymbol{y}$) if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{j} x_{(i)} \geq \sum_{i=1}^{j} y_{(i)}$ for all $j = 1, \ldots, n-1$.

Clearly, $\boldsymbol{x} \overset{\text{m}}{\preceq} \boldsymbol{y}$ implies $\boldsymbol{x} \overset{\text{w}}{\preceq} \boldsymbol{y}$.

The random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to follow the PH model if X_i has the survival function $\bar{G}_i(x) = \dfrac{\alpha_i \bar{F}(x)}{1 - \bar{\alpha}_i \bar{F}(x)}$ for $\alpha_i > 0$, $i = 1, \ldots, n$, where $\bar{F}$ is the baseline survival function and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is the proportional odd vector. Specifically, by $\boldsymbol{X} \sim \text{PO}(\bar{F}, \boldsymbol{\alpha}, \phi)$ we denote the sample having the Archimedean survival copula with generator ϕ and following a PH model with baseline survival function $\bar{F}$ and proportional odd vector $\boldsymbol{\lambda}$.

2 Main result

Now, we state our main results. For the PO samples with Archimedean survival copulas, we present here the usual stochastic order on the sample minimum. The smallest order statistic $X_{1:n}$ of the sample $\boldsymbol{X}$ gets survival function

$$\bar{G}_{X_{1:n}}(x) = \phi\Big(\sum_{i=1}^{n} \psi\Big(\frac{\alpha_i \bar{F}(x)}{1 - \bar{\alpha}_i \bar{F}(x)}\Big)\Big) = J_1(x, \boldsymbol{\alpha}, \phi) \tag{2.1}$$

Theorem 2.1. For $\boldsymbol{X} \sim \text{PO}(\bar{F}, \boldsymbol{\alpha}, \phi_1)$ and $\boldsymbol{X^*} \sim \text{PO}(\bar{F}, \boldsymbol{\alpha^*}, \phi_2)$, if $\psi_2 \circ \phi_1$ is super-additive, then $\boldsymbol{\alpha} \overset{\text{w}}{\succeq} \boldsymbol{\alpha^*}$ implies $X_{1:n} \leq_{\text{st}} X^*_{1:n}$.

Proof. According to Equation (2.1), $X_{1:n}$ and $X^*_{1:n}$ have their respective survival functions $J_1(x, \boldsymbol{\alpha}, \phi_1)$ and $J_1(x, \boldsymbol{\alpha^*}, \phi_2)$, for $x \geq 0$. First we show that $J_1(x, \boldsymbol{\alpha}, \phi_1)$ is increasing and Schur-concave function of $\alpha_i, i = 1, \ldots, n$. Since ϕ_1 is decreasing, we have

$$\frac{\partial J_1(x, \boldsymbol{\alpha}, \phi_1)}{\partial \alpha_i} =$$

$$\frac{\bar{F}(x)(1 - \bar{F}(x))}{(1 - \bar{\alpha}_i \bar{F}(x))^2} \frac{1}{\phi_1'(\psi_1(\dfrac{\alpha_i \bar{F}(x)}{1 - \bar{\alpha}_i \bar{F}(x)}))} \phi_1'\Big(\sum_{i=1}^{n} \psi_1\Big(\frac{\alpha_i \bar{F}(x)}{1 - \bar{\alpha}_i \bar{F}(x)}\Big)\Big) \geq 0,$$

That is, $J_1(x, \boldsymbol{\alpha}, \phi_1)$ is increasing in α_i for $i = 1, \ldots, n$. Furthermore, for $i \neq j$,

$$(\alpha_i - \alpha_j)\Big(\frac{\partial J_1(x, \boldsymbol{\alpha}, \phi_1)}{\partial \alpha_i} - \frac{\partial J_1(x, \boldsymbol{\alpha}, \phi_1)}{\partial \alpha_j}\Big) =$$

$$(\alpha_i - \alpha_j)\bar{F}(x)(1 - \bar{F}(x))\phi_1'\Big(\sum_{i=1}^{n} \psi_1\Big(\frac{\alpha_i \bar{F}(x)}{1 - \bar{\alpha}_i \bar{F}(x)}\Big)\Big)$$

$$\Big(\frac{1}{(1-\bar{\alpha}_i\bar{F}(x))^2\phi_1'(\psi_1(\frac{\alpha_i\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))} - \frac{1}{(1-\bar{\alpha}_j\bar{F}(x))^2\phi_1'(\psi_1(\frac{\alpha_j\bar{F}(x)}{1-\bar{\alpha}_j\bar{F}(x)}))}\Big).$$

Let

$$h(\alpha_i) = (1-\bar{\alpha}_i\bar{F}(x))^2\phi_1'(\psi_1(\frac{\alpha_i\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)})),$$

then, by the decreasing and convexity of ϕ_1, we have

$$\frac{\partial h(\alpha_i)}{\partial\alpha_i} = 2(1-\bar{\alpha}_i\bar{F}(x))\bar{F}(x)\phi_1'(\psi_1(\frac{\alpha_i\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))+$$

$$\bar{F}(x)(1-\bar{F}(x))\frac{\phi_1''(\psi_1(\frac{\alpha_i\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}{\phi_1'(\psi_1(\frac{\alpha_i\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))} \le 0.$$

So, for $i \neq j$,

$$(\alpha_i-\alpha_j)\big(\frac{\partial J_1(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_i} - \frac{\partial J_1(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_j}\big) \le 0.$$

Then Schur-concavity of $J_1(x,\boldsymbol{\alpha},\phi_1)$ follows from Theorem 3.A.4. in [17]. According to Theorem 3.A.8 of [17] $\boldsymbol{\alpha} \overset{\text{w}}{\succeq} \boldsymbol{\alpha}^*$ implies $J_1(x,\boldsymbol{\alpha},\phi_1) \le J_1(x,\boldsymbol{\alpha}^*,\phi_1)$. On the other hand, since $\psi_2\circ\phi_1$ is super-additive by Lemma A.1. of [15], we have $J_1(x,\boldsymbol{\alpha}^*,\phi_1) \le J_1(x,\boldsymbol{\alpha}^*,\phi_2)$. So, it holds that

$$J_1(x,\boldsymbol{\alpha},\phi_1) \le J_1(x,\boldsymbol{\alpha}^*,\phi_1) \le J_1(x,\boldsymbol{\alpha}^*,\phi_2).$$

That is, $X_{1:n} \le_{\text{st}} X^*_{1:n}$. □

In the following Theorem, we study the largest order statistic from PO models with Archimedean copula. The largest order statistic $X_{n:n}$ of the sample $\boldsymbol{X}$ gets distribution function

$$G_{X_{n:n}}(x) = \phi\big(\sum_{i=1}^{n}\psi(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)})\big) = J_2(x,\boldsymbol{\alpha},\phi) \tag{2.2}$$

Theorem 2.2. For $\boldsymbol{X} \sim \text{PO}(\bar{F},\boldsymbol{\alpha},\phi_1)$ and $\boldsymbol{X}^* \sim \text{PO}(\bar{F},\boldsymbol{\alpha}^*,\phi_2)$, if ϕ_1 or ϕ_2 is log-concave, and $\psi_1\circ\phi_2$ is super-additive, then $\boldsymbol{\alpha} \overset{\text{w}}{\succeq} \boldsymbol{\alpha}^*$ implies $X_{n:n} \le_{\text{st}} X^*_{n:n}$.

Proof. According to Equation (2.2), $X_{n:n}$ and $X^*_{n:n}$ have their respective distribution functions $J_2(x,\boldsymbol{\alpha},\phi_1)$ and $J_2(x,\boldsymbol{\alpha}^*,\phi_2)$, for $x \ge 0$. We only prove the case that ϕ_1 is log-concave, and the other case can be finished similarly. First we show that $J_2(x,\boldsymbol{\alpha},\phi_1)$ is decreasing and Schur-convex function of $\alpha_i, i = 1,\ldots,n$. Since ϕ_1 is decreasing, we have

$$\frac{\partial J_2(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_i} =$$

$$\frac{-\bar{F}(x)(1-\bar{F}(x))}{(1-\bar{\alpha}_i\bar{F}(x))^2}\frac{1}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}\phi_1'\big(\sum_{i=1}^{n}\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)})\big)\leq 0,$$

That is, $J_2(x,\boldsymbol{\alpha},\phi_1)$ is decreasing in α_i for $i=1,\ldots,n$. Furthermore, for $i\neq j$,

$$(\alpha_i-\alpha_j)\big(\frac{\partial J_2(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_i}-\frac{\partial J_2(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_j}\big)=$$

$$(\alpha_i-\alpha_j)(-\bar{F}(x))\phi_1'\big(\sum_{i=1}^{n}\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)})\big)$$

$$\left(\frac{1}{1-\bar{\alpha}_i\bar{F}(x)}\frac{\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}-\frac{1}{1-\bar{\alpha}_j\bar{F}(x)}\frac{\frac{1-\bar{F}(x)}{1-\bar{\alpha}_j\bar{F}(x)}}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_j\bar{F}(x)}))}\right).$$

Note that the log-concavity of ϕ_1 implies the increasing property of $\frac{\phi_1}{\phi_1'}$. Since $\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)})$ is increasing in α_i, then $\dfrac{\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}=\dfrac{\phi_1(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_j\bar{F}(x)}))}$ is increasing in α_i. Also $\dfrac{1}{1-\bar{\alpha}_i\bar{F}(x)}$ is a decreasing function of α_i, and thus $\dfrac{1}{1-\bar{\alpha}_i\bar{F}(x)}\dfrac{\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}}{\phi_1'(\psi_1(\frac{1-\bar{F}(x)}{1-\bar{\alpha}_i\bar{F}(x)}))}$ is increasing in α_i. So, for $i\neq j$,

$$(\alpha_i-\alpha_j)\big(\frac{\partial J_2(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_i}-\frac{\partial J_2(x,\boldsymbol{\alpha},\phi_1)}{\partial\alpha_j}\big)\geq 0.$$

Then Schur-convexity of $J_2(x,\boldsymbol{\alpha},\phi_1)$ follows from Theorem 3.A.4. in [17]. According to Theorem 3.A.8 of [17] $\boldsymbol{\alpha}\overset{\text{w}}{\succeq}\boldsymbol{\alpha}^*$ implies $J_2(x,\boldsymbol{\alpha},\phi_1)\geq J_2(x,\boldsymbol{\alpha}^*,\phi_1)$. On the other hand, since $\psi_1\circ\phi_2$ is super-additive by Lemma A.1. of [15], we have $J_2(x,\boldsymbol{\alpha}^*,\phi_1)\geq J_2(x,\boldsymbol{\alpha}^*,\phi_2)$. So, it holds that

$$J_2(x,\boldsymbol{\alpha},\phi_1)\geq J_2(x,\boldsymbol{\alpha}^*,\phi_1)\geq J_2(x,\boldsymbol{\alpha}^*,\phi_2).$$

That is, $X_{n:n}\leq_{\text{st}}X_{n:n}^*$. □

3 Conclusions

In this paper, we studied extreme order statistics from random variables following the proportional odds model. In the presence of the Archimedean copula for the random variables, we obtained new results on the usual stochastic ordering of the smallest and largest order statistics.

References

[1] Balakrishnan, N. and Zhao, P. (2013), Ordering properties of order statistics from heterogeneous populations: a review with an emphasis on some recent developments, *Probability in the Engineering and Informational Sciences*, **27**, 403-443.

[2] Bashkar, E., Torabi, H., Dolati, A. and Belzunce, F. (2018). f-Majorization with Applications to Stochastic Comparison of Extreme Order Statistics. *Journal of Statistical Theory and Applications*, **17(3)**, 520-536.

[3] Bashkar, E., Torabi, H. and Roozegar, R. (2017). Stochastic comparisons of extreme order statistics in the heterogeneous exponentiated scale model. *Journal of Statistical Theory and Applications*, **16(2)**, 219-238.

[4] Bennett, S. (1983). Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**, 273-277.

[5] Boland, P. J., EL-Neweihi, E., and Proschan, F. (1994). Applications of the hazard rate ordering in reliability and order statistics. *Journal of Applied Probability*, **31**, 180–192.

[6] Dolati, A. (2008). On Dependence Properties of Random Minima and Maxima. *Communications in Statistics - Theory and Methods*, **38**, 393–399.

[7] Dolati, A., Genest, C. and Kochar, S.C. (2008). On the dependence between the extreme order statistics in the proportional hazards model. *Journal of Multivariate Analysis*, **99**, 777–786.

[8] Dykstra, R., Kochar, S. and Rojo, J. (1997). Stochastic comparisons of parallel systems of heterogeneous exponential components. *Journal of Statistical Planning and Inference*, **69**, 203–211.

[9] Fang, L. and Balakrishnan, N. (2016). Likelihood ratio order of parallel systems with heterogeneous Weibull components. *Metrika*, **79**, 693-703.

[10] Fang, R., Li, C. and Li, X. (2015). Stochastic comparisons on sample extremes of dependent and heterogenous observations. *Statistics*, 1-26.

[11] Fang, L. and Zhang, X. (2013). Stochastic comparisons of series systems with heterogeneous Weibull components. *Statistics and Probability Letters*, **83**, 1649-1653.

[12] Fang, L. and Zhang, X. (2015). Stochastic comparisons of parallel systems with exponentiated Weibull components. *Statistics and Probability Letters*, **97**, 25-31.

[13] Khaledi, B. E., Farsinezhad, S. and Kochar, S. C. (2011). Stochastic comparisons of order statistics in the scale model. *Journal of Statistical Planning and Inference*, **141(1)**, 276-286.

[14] Khaledi, B.E., Kochar, S.C. (2006). Weibull distribution: some stochastic comparisons results. *Journal of Statistical Planning and Inference*, **136**, 3121-3129.

[15] Li, X. and Fang, R. (2015). Ordering properties of order statistics from random variables of Archimedean copulas with applications, *Journal of Multivariate Analysis*, **133**, 304-320.

[16] Li, C. and Li, X. (2015). Likelihood ratio order of sample minimum from heterogeneous Weibull random variables. *Statistics and Probability Letters*, **97**, 46-53.

[17] Marshall, A.W., Olkin, I. and Arnold, B.C. (2011). *Inequalities: Theory of Majorization and its Applications*. Springer, New York.

[18] McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean Copulas, d-Monotone Functions and ℓ_1-Norm Symmetric Distributions, *The Annals of Statistics*, 3059-3097.

[19] Nelsen, R. B. (2006). *An introduction to copulas*. Springer, New York.

[20] Pledger, G. and Proschan, F. (1971). Comparisons of order statistics and of spacings from heterogeneous distributions. In: Rustagi, J.S. (Ed.), *Optimizing Methods in Statistics*. Academic Press, New York, pp. 89-113.

[21] Proschan, F. and Sethuraman, J. (1976). Stochastic comparisons of order statistics from heterogeneous populations, with applications in reliability. *Journal of Multivariate Analysis*, **6**, 608–616.

[22] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*. Springer, New York.

[23] Torrado, N. (2017). Stochastic comparisons between extreme order statistics from scale models. Statistics, **51(6)**, 1359-1376.

[24] Zhao, P., Li, X. and Balakrishnan, N. (2009). Likelihood ratio order of the second order statistic from independent heterogeneous exponential random variables. *Journal of Multivariate Analysis*, **100**, 952-962.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Characterizations of Circular−Circular Copula Using Extended Copulas

Hatami, M. [1] Alamatsaz, M. H. [2]

Department of Statistics, University of Isfahan, Isfahan, Iran

Abstract

Joining marginal circular distribution functions by copulas does not necessarily lead to joint circular distribution functions. In this paper, considering an extended notion of copulas, we propose a new classe of bivariate copulas, called circular-circular copulas, in order to construct bivariate circular distributions with known univariate circular marginals. We shall provide new definition and some novel characterizations for such extended copulas and study their useful properties. We, then, describe a method for constructing circular-circular copulas. Nonnegative trigonometric sums (NNTS) copula is introduced and its related measures of dependence are obtained. Finally, we shall develop a method of generalizing circular-circular copulas to multivariate circular copulas.

Keywords: Nonnegative trigonometric sums, Bivariate circular distribution, Multivariate circular copula.

[1] m.hatami.v@gmail.com
[2] alamatho@sci.ui.ac.ir

1 Introduction

Copula modeling is a useful and popular method in many areas of Statistics such as in Finance, Hydrology, Drought study, among many others. Copulas are of interest to statisticians for two main reasons: first as a way of studying scale-free measures of dependence, and second as a starting point for constructing families of bivariate (and more generally, multivariate) distributions.

Copulas do not necessarily lead to a joint circular distribution function when linking marginal circular distribution functions. The main concern of this paper is thus to introduce classes of extended copulas that preserve the circular characteristics by the resulting joint distribution functions.

The paper is organized as follows. In Section 2 we recall definitions of circular distribution functions and standard (linear) copulas. We, then, introduce our classe of circular-circular copulas and study their general characteristics in Section 3. Some novel and interesting characterizations are also provided for such extended copulas in this sections. In Section 4 we shall reveal a method of constructing circular-circular copulas and, in particular, we shall introduce nonnegative trigonometric sums (NNTS) copulas. In Section 5 we obtain related dependence measures of the NNTS copula. In Section 6 we shall develop a method for constructing multivariate circular copulas from circular-circular copulas.

2 Circular distribution function and copula

The circular density of a univariate absolutely continuous circular random variable, Θ, defined on the unit circle, $\mathbb{S}^1$, is a function $f(\theta)$ satisfying the conditions:

1. $f(\theta) \geq 0$ for $-\infty < \theta < \infty$,
2. $f(\theta + 2k\pi) = f(\theta)$ for $k \in \mathbb{Z}, -\infty < \theta < \infty$,
3. $\int_0^{2\pi} f(\theta)d\theta = 1$.

Let $F(\theta) = \int_0^{\theta} f(w)dw$. A circular distribution function (df) is defined by F restricted on $[0, 2\pi]$, i.e.,

$$P(0 < \Theta \leq \theta) = F(\theta), \ \ 0 \leq \theta \leq 2\pi, \tag{2.1}$$

such that

$$F(\theta + 2\pi) - F(\theta) = 1, \ \ -\infty < \theta < \infty. \tag{2.2}$$

As we observe, the circular df F defined above differs from a linear df in having the following mathematical properties:

$$\lim_{\theta \to -\infty} F(\theta) = -\infty, \ \lim_{\theta \to \infty} F(\theta) = \infty.$$

By definition, $F(0) = 0,\ F(2\pi) = 1$ ([7]). Eq. (2.2) can be generalized as

$$F(\theta + 2k\pi) - F(\theta) = k,\ k \in \mathbb{Z},\ -\infty < \theta < \infty. \tag{2.3}$$

In general, then, the values taken by the mathematical form of a circular df, i.e. F, are clearly not probabilities. Similar to the univariate case, the joint distribution function of a bivariate circular random vector must satisfy certain properties. The density function of a bivariate circular random vector is defined as follows:

Definition 1. *Let f be a bivariate function defined on the surface of a torus, $\mathbb{S}^1 \times \mathbb{S}^1$, such that*

1. $f(\theta_1, \theta_2) \geq 0\ \ for\ \ -\infty < \theta_1, \theta_2 < \infty$,
2. $f(\theta_1 + 2k\pi, \theta_2 + 2j\pi) = f(\theta_1, \theta_2)\ \ for\ k, j \in \mathbb{Z}, -\infty < \theta_1, \theta_2 < \infty$,
3. $\int_0^{2\pi} \int_0^{2\pi} f(\theta_1, \theta_2) d\theta_2 d\theta_1 = 1$.

Then, $f(\theta_1, \theta_2)$ is said to be the bivariate circular probability density function (pdf) of a circular random vector (Θ_1, Θ_2).

Let

$$F(\theta_1, \theta_2) = \int_0^{\theta_1} \int_0^{\theta_2} f(w, z) dz dw.$$

A bivariate circular df is defined by F restricted on $[0, 2\pi] \times [0, 2\pi]$, i.e.,

$$P(0 \leq \Theta_1 \leq \theta_1, 0 \leq \Theta_2 \leq \theta_2) = F(\theta_1, \theta_2),\ \ 0 \leq \theta_1 \leq\ 2\pi,\ 0 \leq \theta_2 \leq\ 2\pi. \tag{2.4}$$

It is easily observed that we have

$$F(\theta_1 + 2\pi, \theta_2 + 2\pi) = 1 + F_1(\theta_1) + F_2(\theta_2) + F(\theta_1, \theta_2),$$

and thus

$$F(\theta_1 + 2\pi, \theta_2 + 2\pi) - F(\theta_1, \theta_2) = 1 + F_1(\theta_1) + F_2(\theta_2),\ -\infty < \theta_1, \theta_2 < \infty, \tag{2.5}$$

where f_1, f_2 stand for the marginal densities and F_1, F_2 for the associated dfs. We have used here the fact that

$$F(\theta_1 + 2\pi, \theta_2) = F_2(\theta_2) + F(\theta_1, \theta_2),$$

which we can summarize as

$$F(\theta_1 + 2\pi, \theta_2) - F(\theta_1, \theta_2) = F_2(\theta_2),\ -\infty < \theta_1, \theta_2 < \infty. \tag{2.6}$$

Similarly, we obtain $F(\theta_1, \theta_2 + 2\pi) - F(\theta_1, \theta_2) = F_1(\theta_1)$, $-\infty < \theta_1, \theta_2 < \infty$.

It is necessary to note that the marginal and joint distribution functions $F_1(\theta_1)$, $F_2(\theta_2)$ and $F(\theta_1, \theta_2)$ are mathematically defined on $\mathbb{R}$ and $\mathbb{R}^2$, respectively, and they are circular distribution functions on the restricted intervals $0 \leq \theta_1, \theta_2 \leq 2\pi$. Due to the periodicity of their density functions, these functions possess certain unique characteristics such as those in (2.2)–(2.6). In what follows, we make use of property (2.5) to check whether a bivariate joint df is circular or not.

Lemma 2.1. *Let $F_1(\theta_1)$ and $F_2(\theta_2)$ be circular marginal df's and $F(\theta_1, \theta_2)$ be their bivariate joint df on $0 \leq \theta_1, \theta_2 \leq 2\pi$ such that for $-\infty < \theta_1, \theta_2 < \infty$, $F(\theta_1 + 2\pi, \theta_2 + 2\pi) - F(\theta_1, \theta_2) = 1 + F_1(\theta_1) + F_2(\theta_2)$. Then, $F(\theta_1, \theta_2)$ is a bivariate circular df.*

3 Circular-circular copulas and their general properties

Let C be a copula with density c. Then, the joint density function of circular marginals F_1 and F_2 is $f(\theta_1, \theta_2) = f_1(\theta_1) f_2(\theta_2) c(F_1(\theta_1), F_2(\theta_2))$. This density is a bivariate circular density function if it satisfies condition 2 of Definition 1. But, due to the fact that periodicity of circular distributions, here $F_1(\theta_1 + 2\pi)$ and $F_2(\theta_2 + 2\pi)$ are not restricted to [0, 1], we cannot deal with periodicity of the density function f. To overcome this situation, we have to extend the notion of copula over $\mathbb{R}^2$. Thus, we introduce the notion of extended copulas. We call a bivariate function $C(u, v) : \mathbb{R}^2 \to \mathbb{R}$ extended copula if its restriction to $[0, 1] \times [0, 1]$ is a standard (linear) copula. For instance, $C(u, v) = uv : \mathbb{R}^2 \to \mathbb{R}$ is an extended copula because $uv : [0, 1]^2 \to [0, 1]$ is the standard independent copula.

Now, the question is that, is there at least one extended copula C such that $C(F_1(\theta_1), F_2(\theta_2))$ leads to a joint circular distribution function?
To answer this question, consider the extended copula $C(u, v) = uv : \mathbb{R}^2 \to \mathbb{R}$. We have $C(F_1(\theta_1), F_2(\theta_2)) = F_1(\theta_1) F_2(\theta_2)$. Since $0 \leq F_i(\theta_i) \leq 1$ for $0 \leq \theta_i \leq 2\pi; i = 1, 2$; here extended copula is an standard copula for which Sklar's Theorem holds. Therefore, $F(\theta_1, \theta_2) = C(F_1(\theta_1), F_2(\theta_2)) = F_1(\theta_1) F_2(\theta_2), 0 \leq \theta_1, \theta_2 \leq 2\pi$, is a distribution function. Also, for $-\infty < \theta_1, \theta_2 < \infty$, $F(\theta_1 + 2\pi, \theta_2 + 2\pi) - F(\theta_1, \theta_2) = 1 + F_1(\theta_1) + F_2(\theta_2)$. Thus, $C(F_1(\theta_1), F_2(\theta_2)) = F_1(\theta_1) F_2(\theta_2)$ satisfies Lemma 2.1 and hence it is a joint circular distribution function. Therefore, the answer is positive, i.e., the set of such extended copulas is not empty. In this section, we develop a large class of such extended copulas and discuss their properties.

Definition 2. *Let F_1 and F_2 be two univariate circular distribution functions and $\mathfrak{C}_{cc}$ be a class of extended copulas such that for any $C \in \mathfrak{C}_{cc}$, $F(\theta_1, \theta_2) = C(F_1(\theta_1), F_2(\theta_2))$ is a joint circular distribution function. Then, we call $\mathfrak{C}_{cc}$ the class of circular-circular (cc) copulas.*

In the following, we give a useful characterization for the members of the set $\mathfrak{C}_{cc}$.

Theorem 3.1. *Let C be an absolutely continuous extended copula. Then $C \in \mathfrak{C}_{cc}$ if, and only if, the extended copula density, c, is a 1-periodic function.*

According to Theorem 3.1, we can define cc copula densities as follows. A function c is the bivariate probability density function of an absolutely continuous cc copula if, and only if,

1. $c(u,v) \geq 0 \;\; for \;\; -\infty < u,v < \infty$,
2. $c(u+k, v+j) = c(u,v) \;\; for \; k,j \in \mathbb{Z}, -\infty < u,v < \infty$,
3. $\int_0^1 \int_0^1 c(u,v)dudv = 1$,
4. $c(u) = \int_0^1 c(u,v)dv = 1, \;\; c(v) = \int_0^1 c(u,v)du = 1$.

Let

$$C(u,v) = \int_0^u \int_0^v c(w,z)dzdw.$$

Then, cc copula, i.e., the bivariate df corresponding to c, is defined by C restricted on $[0,1] \times [0,1]$, i.e.,

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

It is easily observed that we have

$$C(1+u, 1+v) = 1 + u + v + C(u,v),$$

and thus,

$$C(1+u, 1+v) - C(u,v) = 1 + u + v, \quad -\infty < u,v < \infty. \tag{3.1}$$

Further, we have

$$\begin{aligned} C(1+u, v) &= v + C(u,v), \quad -\infty < u,v < \infty, \\ C(u, 1+v) &= u + C(u,v), \quad -\infty < u,v < \infty. \end{aligned} \tag{3.2}$$

Due to the periodicity of $c(u,v)$, the bivariate function $C(u,v)$ can be defined on $\mathbb{R}^2$. Therefore, $C(u,v) : \mathbb{R}^2 \to \mathbb{R}$ can be seen as a extended copula whose restriction to $[0,1]\times[0,1]$ is the standard (linear) copula. Here, we present a new definition for cc copulas.

Definition 3. *A bivariate function $C(u,v)$ is a cc copula, i.e., $C \in \mathfrak{C}_{cc}$, if, it satisfies the following two conditions*

1. *$C(u,v)$ is a copula for $[u,v] \in [0,1] \times [0,1]$,*

2. $C(1+u,1+v)-C(u,v)=1+u+v$ *for* $[u,v]\in\mathbb{R}\times\mathbb{R}$.

We also have the following characterization for certain subclass of $\mathfrak{C}_{cc}$. De la Pena [1] showed that, any absolutely continuous copula can be represented by the form $C(u,v)=uv+A(u,v)$. However, determining the function $A(u,v)$ for a given copula is not an easy task. In the following theorem, we show that cc copula can be represented as $C(u,v)=uv+A(u,v)$. Conditions on $A(u,v)$ for a cc copula are specifically mentioned in Theorem 3.2. We can also use the proof of Theorem 3.2 to find $A(u,v)$ uniquely for any cc copula.

Theorem 3.2. *Let* C *be absolutely continuous. Then,* $C\in\mathfrak{C}_{cc}$ *if, and only if,* C *has the form*

$$C(u,v)=uv+A(u,v), \tag{3.3}$$

where $A(u,v)$ *is an absolutely continuous function satisfying the following conditions:*

a) $A(u,0)=A(0,v)=A(u,1)=A(1,v)=0,\ \ for\ \ -\infty<u,v<\infty$,

b) $A(u_1,v_1)-A(u_1,v_2)-A(u_2,v_1)+A(u_2,v_2)\geq(u_2-u_1)(v_1-v_2)$ *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) $A(u+1,v+1)=A(u,v),\ \ for\ \ -\infty<u,v<\infty$.

4 Construction of cc copulas

The main problem in constructing parametric extended copulas of class $\mathfrak{C}_{cc}$ is finding the function $A(u,v)$, in Eq. (1.6), satisfying the conditions of Theorem 3.2.

Here, we shall introduce an approach to construct cc copulas by using truncated Fourier series, called trigonometric polynomials. A method for modeling univariate, bivariate circular and spherical data based on nonnegative trigonometric sums is applied by [2]. A nonnegative trigonometric sums is a partial sum of the terms in a Fourier series. We are looking for conditions that makes the truncated Fourier series

$$\begin{aligned} f(u,v)=1&+4\sum_{n=1}^{k}\sum_{m=1}^{j}\alpha_{n,m}\cos(2\pi nu)\cos(2\pi mv)\\ &+4\sum_{n=1}^{k}\sum_{m=1}^{j}\beta_{n,m}\cos(2\pi nu)\sin(2\pi mv)\\ &+4\sum_{n=1}^{k}\sum_{m=1}^{j}\gamma_{n,m}\sin(2\pi nu)\cos(2\pi mv)\\ &+4\sum_{n=1}^{k}\sum_{m=1}^{j}\delta_{n,m}\sin(2\pi nu)\sin(2\pi mv), \end{aligned} \tag{4.1}$$

nonnegative. For this propose, we express the truncated Fourier series (4.1) as the squared norm of a sum of complex numbers, i.e.,

$$\begin{aligned} f(u,v) &= \left\| \sum_{n=0}^{k} \sum_{m=0}^{j} p_{n,m} e^{i2\pi(nu+mv)} \right\|^2 \\ &= \sum_{n_1=0}^{k} \sum_{m_1=0}^{j} \sum_{n_2=0}^{k} \sum_{m_2=0}^{j} p_{n_1,m_1} \bar{p}_{n_2,m_2} e^{i2\pi[(n_1-n_2)u+(m_1-m_2)v]}, \end{aligned} \tag{4.2}$$

such that

$$\sum_{n=0}^{k} \sum_{m=0}^{j} |p_{n,m}|^2 = 1, \quad \sum_{\nu=0}^{k-n} \sum_{\eta=0}^{j} p_{\nu+n,\eta} \bar{p}_{\nu,\eta} = 0, \quad \sum_{\nu=0}^{k} \sum_{\eta=0}^{j-m} p_{\nu,\eta+m} \bar{p}_{\nu,\eta} = 0,$$

where $p_{n,m}$, $n = 0, 1, ..., k$, $m = 0, 1, ..., j$, are complex numbers and $\bar{p}_{n,m}$ refers to the complex conjugate of $p_{n,m}$. After extending (4.2), we have

$$\begin{aligned} f(u,v) = 1 &+ 4 \sum_{n=1}^{k} \sum_{m=1}^{j} a_{n,m} \cos(2\pi nu + 2\pi mv) + b_{n,m} \sin(2\pi nu + 2\pi mv) \\ &+ 4 \sum_{n=1}^{k} \sum_{m=1}^{j} c_{n,m} \cos(2\pi nu - 2\pi mv) + d_{n,m} \sin(2\pi nu - 2\pi mv). \end{aligned} \tag{4.3}$$

Thus, the bivariate trigonometric polynomial of order (k, j) in (4.3) is nonnegative for every real u and v if there exist complex numbers $p_{n,m}, n = 1, ..., k, m = 1, ..., j$, such that

$$a_{n,m} - ib_{n,m} = 2 \sum_{\nu=0}^{k-n} \sum_{\eta=0}^{j-m} p_{\nu+n,\eta+m} \bar{p}_{\nu,\eta}, \quad c_{n,m} - id_{n,m} = 2 \sum_{\nu=0}^{k-n} \sum_{\eta=m}^{j} p_{\nu+n,\eta-m} \bar{p}_{\nu,\eta},$$

for $n = 1, ..., k, m = 1, ..., j$. Consequently, the function (4.3) is a density function where we can be rewritten as (4.1). Addition to this distribution having uniform marginals, (4.3) is a cc copula density. The trigonometric moments of this distribution are easily expressed in terms of the parameters of the distribution. Indeed, given the orthogonality properties of the terms of a Fourier series, we have:

$$\begin{aligned} \alpha_{n,m} &= E(\cos(2\pi nu)\cos(2\pi mv)) = a_{n,m} + c_{n,m}, \\ \beta_{n,m} &= E(\cos(2\pi nu)\sin(2\pi mv)) = b_{n,m} - d_{n,m}, \\ \gamma_{n,m} &= E(\sin(2\pi nu)\cos(2\pi mv)) = b_{n,m} + d_{n,m}, \\ \delta_{n,m} &= E(\sin(2\pi nu)\sin(2\pi mv)) = c_{n,m} - a_{n,m}. \end{aligned}$$

Thus, nonnegative trigonometric sums (NNTS) copula is obtained as

$$\begin{aligned} C(u,v) = uv &+ 4\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{a_{n,m}+c_{n,m}}{4\pi^2 mn}\sin(2\pi nu)\sin(2\pi mv) \\ &+ 4\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{b_{n,m}-d_{n,m}}{4\pi^2 mn}\sin(2\pi nu)(1-\cos(2\pi mv)) \\ &+ 4\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{b_{n,m}+d_{n,m}}{4\pi^2 mn}(1-\cos(2\pi nu))\sin(2\pi mv) \\ &+ 4\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{c_{n,m}-a_{n,m}}{4\pi^2 mn}(\cos(2\pi nu)-1)(\cos(2\pi mv)-1), \\ = uv + A(u,v), &\quad \forall\ n=1,...,k, m=1,...,j. \end{aligned}$$

Note that the NNTS copula for the special case $a_{n,m} = b_{n,m} = c_{n,m} = d_{n,m} = 0$, $n = 1,...,k$ and $m = 1,...,j$, coincides with the independent copula. T

5 Measures of association

In this section, we consider measures of association for either circular data or usual linear (non circular) data, because the NNTS copula can be used circular as well as linear distribution functions. We calculate four dependence measures for circular data and three measures of association for linear data.

Theorem 5.1. *The values of Kendall's tau, Spearman's rho and Gini's gamma associated with NNTS copula, C, are, respectively, given by*

$$\tau_C = \frac{8}{\pi^2}\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{1}{mn}(\delta_{n,m}+\delta_{n,m}\alpha_{n,m}-\gamma_{n,m}\beta_{n,m}),$$

$$\rho_C = \frac{3}{\pi^2}\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{1}{mn}\delta_{n,m}$$

and

$$\varphi_C = \frac{3}{\pi^2}\left[\sum_{m=1}^{j}\frac{1}{m^2}\alpha_{m,m}+\sum_{n=1}^{k}\sum_{m=1}^{j}\frac{3}{n}\delta_{n,m}\right].$$

We consider four dependence measures for circular data, namely those of [3,4,5,6].

Theorem 5.2. *The values of dependence measures of Fisher and Lee (ρ_{FL}), Jammalamadaka and Sarma (ρ_{JS}), Johnson and Wehrly (ρ_{JW}) and Jupp and Mardia (ρ_{JM}), for the circula density of NNTS are, respectively, given by* $\rho_{FL} = 2(\alpha_{1,1}\delta_{1,1} + \beta_{1,1}\gamma_{1,1})$, $\rho_{JS} = \sqrt{\rho_1} - \sqrt{\rho_2}$,

$$\rho_{JW} = \sqrt{1/2}\left(\alpha_{1,1}^2 + \beta_{1,1}^2 + \gamma_{1,1}^2 + \delta_{1,1}^2 + \sqrt{\rho_1\rho_2}\right)^{1/2}$$

and

$$\rho_{JM} = 4(\alpha_{1,1}^2 + \beta_{1,1}^2 + \gamma_{1,1}^2 + \delta_{1,1}^2),$$

where $\rho_1 = (\alpha_{1,1} + \delta_{1,1})^2 + (\gamma_{1,1} - \beta_{1,1})^2$ *and* $\rho_2 = (\alpha_{1,1} - \delta_{1,1})^2 + (\gamma_{1,1} + \beta_{1,1})^2$.

6 A multivariate generalization

Here, we present a method for constructing a multivariate circular copula (circular copulas of dimension $n > 2$). In this method, multivariate circular copulas are obtained directly in terms of cc copulas.

We can obtain results similar to Theorem 3.1 and Definition 3 for multivariate circular copulas.

Theorem 6.1. *Let C be an absolutely continuous multivariate extended copula. Then, C is a multivariate circular copula if, and only if, the multivariate extended copula density, c, is a 1-periodic function.*

According to Theorem 6.1, we may define multivariate circular copula densities as follows. A function c is the multivariate probability density function of an absolutely continuous multivariate circular copula if, and only if,

1. $c(u_1, ..., u_n) \geq 0$ *for* $-\infty < u_1, ..., u_n < \infty$,
2. $c(u_1 + k_1, ..., u_n + k_n) = c(u_1, ..., u_n)$ *for* $k_1, ..., k_n \in \mathbb{Z}, -\infty < u_1, ..., u_n < \infty$,
3. $\int_0^1 ... \int_0^1 c(u_1, ..., u_n)du_1...du_n = 1$,
4. $c(u_i) = \int_0^1 c(u_1, ..., u_n)du_i = 1,\ i = 1, ...n.$

We may, then, define multivariate cc copulas as below.

Definition 4. *A multivariate function $C(u_1, ..., u_n)$ is a multivariate circular copula if it satisfies the following two conditions*

1. *$C(u_1, ..., u_n)$ be a multivariate copula for $[u_1, ..., u_n] \in [0, 1] \times ... \times [0, 1]$.*

2.

$$\begin{aligned} C(u_1+1,...,u_n+1) =& 1+\sum_{i=1}^{n} u_i + \sum_{1\le j_1<j_2\le n} C_{j_1,j_2}(u_{j_1},u_{j_2}) \\ &+ ... + \sum_{1\le j_1<...<j_{n-1}\le n} C_{j_1,...,j_{n-1}}(u_{j_1},...,u_{j_{n-1}}), \\ &+ C(u_1,...,u_n), \quad for\ [u_1,...,u_n]\in \mathbb{R}\times ...\times \mathbb{R}. \end{aligned}$$

Characterization of multivariate circular copulas is more complicated than cc copulas. In what follows, we obtain a multivariate circular copula using cc copulas.

Theorem 6.2. *Let $C_1,...,C_n$ be n cc copulas. Then,*

$$C(u_1,...,u_n)=\int_0^1 \frac{\partial}{\partial t}C_1(u_1,t).\frac{\partial}{\partial t}C_2(u_2,t)....\frac{\partial}{\partial t}C_n(u_n,t)dt$$

is a multivariate circular copula of the form

$$\begin{aligned} C(u_1,...,u_n) =& \prod_{i=1}^{n} u_i + \sum_{1\le j_1\le n} g_{j_1}(u_{j_1}) \prod_{\substack{i=1\\ i\neq j_1}}^{n} u_i + \sum_{1\le j_1<j_2\le n} g_{j_1,j_2}(u_{j_1},u_{j_2}) \prod_{\substack{i=1\\ i\neq j_1,j_2}}^{n} u_i \\ &+ ... + \sum_{1\le j_1<...<j_{n-1}\le n} g_{j_1,...,j_{n-1}}(u_{j_1},...,u_{j_{n-1}}) \prod_{\substack{i=1\\ i\neq j_1,...,j_{n-1}}}^{n} u_i + g_{1,...,n}(u_1,...,u_n), \end{aligned}$$

where $g_{1,...,k}(u_1,...,u_k), \quad k=1,...,n$, are 1-periodic k-variate functions.

Note that, using the condition **(a)** of Theorem 3.2 in the bivariate case, the copula's form reduces to

$$C(u_1,u_2)=C(u_1,u_2,1,...,1)=u_1u_2+g_{1,2}(u_1,u_2).$$

References

[1] De la Pena, V. H., Ibragimov, R. and Sharakhmetov, S. (2006). Characterizations of joint distributions, copulas, information dependence and decoupling, with applications to time series. *In Optimality, Institute of Mathematical Statistics*, 183-209.

[2] Fernández-Durán, J. J. and Gregorio-Domínguez, M. M. (2014). Modeling angles in proteins and circular genomes using multivariate angular distributions based on multiple nonnegative trigonometric sums. *Statistical applications in genetics and molecular biology*, **13(1)**, 1-18.

[3] Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika*, **70**, 327-332.

[4] Jammalamadaka, S. R. and Sarma, Y. R. (1988). A correlation coefficient for angular variables. *Statistical theory and data analysis II*, 349-364.

[5] Johnson, R. A. and Wehrly, T. (1977). Measures and models for angular correlation and angular-linear correlation. *Journal of the Royal Statistical Society Series B*, **39**, 222-229.

[6] Jupp, P. E. and Mardia, K. V. (1980). A general correlation coefficient for directional data and related regression problems. *Biometrika*, **67**, 163-173.

[7] Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley, Chichester.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Analysis of Dependent Risk Models based on Sarmanov Copula

Hussam Ahmad [1] Amini, M. [2] Sadeghpour Gildeh, B. [3]

Department of Statistics, Ferdowsi University of Mashhad, Iran

Abstract

This paper extend the compound Poisson risk model to consider the distribution of the maximum surplus before failure when the claim amounts and claim inter-arrival times are depended via a Sarmanov copula. We obtain integro-differential equation for this distribution which satisfies integro-differential equation in the state of independence and dependence via Farlie-Gumbel-Morgenstern (FGM) copula.

Keywords: Risk Models, Sarmanov Copula, Distribution of the Maximum Surplus.

1 Introduction

In the recent researches, within the statistical world, modern risk management techniques have an essential role [10]. The collective risk theory which has been studied in different

[1] ahmadhosam116@gmail.com
[2] m-amini@um.ac.ir
[3] sadeghpour@um.ac.ir

models of the risk business company is one of the famous problems of nonlife insurance field. By presenting an insurance risk model, we have to study the failure probability, i.e., the probability that the risk business ever will below some specified value, thus take risk control of the nonlife business.

Mostly, insurance has been built with had been taking in consideration of independence and the law of large numbers has governed the determination of premiums, the incremental complexity of insurance and reinsurance products has led to the increased actuarial interest in the modeling of dependent risk [3]. Classic risk models depend on an assumption of independence among the claim amounts and the interclaim times. This hypothesis simplifies the study of many quantities under such a framework, however, it has proven to be inadequate and too restrictive in many cases. For example, in Nikoloulopoulos and Karlis [11], they point out that on the occurrence of a catastrophe, the total claim amount and the time elapsed since the previous catastrophe are dependent.

Albrecher and Teugels [1] relax the independence assumptions by introducing an arbitrary dependence structure via a copula for the interclaim time and the subsequent claim size.

Recently, some risk models that allow for specific dependence between the claim amounts and the inter-claim times have been studied, for example in [5].

Cossette et al [4] extended the classical compound Poisson risk model to a dependence structure in which the claim amounts and claim inter-arrival times are dependent but a FarlieGumbelMorgenstern (FGM) copula, and Cossette et al [3] supposed the dependence model via a generalized FGM copula.

The aim of this paper is to analyze the probability of the maximum surplus before failure in a risk model which has dependence structure between claim sizes and inter-claim times via Sarmanov copula which is a general case of FGM copula that is studied in previous papers. We extend the classical compound Poisson risk model to assume the distribution of the maximum surplus before failure where the claim sizes depend on inter-claim times via the Sarmanov copula.

In this research which will be introduced below, we have used a dependence structure among the claim amounts and the interclaim times determined with the Sarmanov copula to the compound Poisson risk model. In our dependent model, we have studied the probability of the maximum surplus before failure and obtain the integro-differential equation for it.

In the statistical papers, many writers (see e.g. Lin & Sendova [9], Cheung & Landriault [2] and Jiang et al. [6]) examined the risk models under the hypotheses that claim sizes are independent of the inter-claim times.

With the help of these hypotheses, various equations can be explicitly calculated for certain classes of claim size distributions such as joint and marginal distributions of failure time, the surplus immediately before failure, the deficit at failure and the claim size causing failure.

Some of the risk models that allow for a specific dependence between claim sizes and inter-claim times are considered. For example, Zhang & Yang [12] examined the GerberShiu function in the compound Poisson risk model by propagation with the dependence between

claim sizes and inter-claim times.

2 The Sarmanov families

Let (X, Y) be a bivariate absolutely continuous random variable with the distribution function

$$H_{X,Y}(x, y) = F(x)G(y)\{1 + \theta\phi_1(F(x))\phi_2(G(y))\}, \quad x, y \in \mathbb{R}, \tag{2.1}$$

such that $E[\phi_1(X)] = E[\phi_2(Y)] = 0$ and $1 + \theta\phi_1(x)\phi_2(y) \geq 0$, and the kernels $\phi_1(x), \phi_2(y)$ are differentiable functions on unit interval so that $H(x, y)$ becomes a distribution function with absolutely continuous marginals $F(x)$ and $G(y)$, and in this case Sarmanov copula will be

$$C(u, v) = uv(1 + \theta\phi_1(u)\phi_2(v)), \quad 0 \leq u, v \leq 1. \tag{2.2}$$

If $\phi_1(t) = \phi_2(t) = 1 - t$, then we have the FGM copula.

3 The dependent risk model

Consider the surplus process

$$R(t) = u + ct - \sum_{i=1}^{N(t)} X_i, \quad t \geq 0, \tag{3.1}$$

where $u \geq 0$ is the initial surplus, c represents the insurer's premium income per unit time. Let $\{X_1, X_2, ...\}$ be independent and identically distributed (i.i.d.) positive random variables representing the successive individual claim amounts. These random variables, identically distributed as the canonical random variable X, are hypothesized to have a common cumulative distribution function $F(x), x \geq 0$, with probability density function $f(x)$, of which the Laplace transform is $\tilde{f}(s) = \int_0^\infty e^{-sx} f(x)dx < \infty$.
The counting process $N(t)$ is a renewal process which is the number of claims up to time t and is assigned as $N(t) = \sup\{n : W_1 + W_2 + ... + W_n \leq t\}$ where the inter-claim times $\{W_i\}_{i=1}$ are assumed (i.i.d) as the canonical r.v. W, have exponential distribution $K(t) = 1 - e^{-\lambda t}$, $t \geq 0$, and its Laplace transform $\tilde{k}(s) = \frac{\lambda}{\lambda+s}$.
In addition, we suppose that $\{(X_i, W_i), i \in \mathbb{N}^+\}$ forms a sequence of i.i.d. random vectors distributed as the canonical r.v. (X, W), in which the components may be dependent. Now, we use the Sarmanov copula to define the dependence between the claim size and the inter-claim time. We have from (2.1)

$$F_{X,W}(x, t) = F(x)K(t)\left(1 + \theta\phi_1(F(x))\phi_2(K(t))\right),$$

and the joint p.d.f of (X, W) is given by

$$f_{X,W}(x,t) = \lambda e^{-\lambda t} + \theta g(t).h(x), \tag{3.2}$$

where

$$h(x) = f(x)\phi_1(F(x)) + F(x)\acute{\phi_1}(F(x)),$$

and

$$g(t) = \lambda e^{-\lambda t}\phi_2(K(t)) + (1 - e^{-\lambda t})\acute{\phi_2}(K(t)).$$

Let T denote the time to failure, then the probability of ultimate failure with initial surplus u is defined by

$$\psi(u) = P\left(T < \infty | R(0) = u\right).$$

4 The distribution of the maximum surplus before failure

For $0 \leq u \leq b$, let

$$G(u,b) = P\left(\sup_{0\leq t\leq T} R(t) < b, T < \infty | R(0) = u\right),$$

denote the probability that failure occurs without the surplus process (3.1) reaching level b prior to failure, obviously, $G(u,b) = 0$ for $b \leq u$.
Note that $G(u,b)$ maybe viewed as the distribution of the maximum surplus before failure. In the following, we derive that $G(u,b)$ satisfies an integro-differential equation, let $\mathcal{I}$ and $\mathcal{D}$ represent the identity and the differentiation operators with respect to (w.r.t) u.

Theorem 4.1. *For $0 \leq u \leq b$, the probability $G(u,b)$ satisfies the Integro-differential equation*

$$\left(-\mathcal{D}^2 + \frac{\lambda^2}{c^2}\mathcal{I}\right) G(u,b) = \left(\frac{\lambda^2}{c^2}\mathcal{I} + \frac{\lambda}{c}\mathcal{D}\right)\gamma_1(u) + \theta\left(\frac{\lambda^2}{c^2}\mathcal{I} - \mathcal{D}^2\right) M(u), \tag{4.1}$$

where

$$\gamma_1(u) = \int_0^u G(u-x,b)f(x)dx + \int_u^\infty f(x)dx, \tag{4.2}$$

$$\gamma_2(u) = \int_0^u G(u-x,b)h(x)dx + \int_u^\infty h(x)dx, \tag{4.3}$$

and

$$M(u) = \int_0^{\frac{b-u}{c}} g(t)\gamma_2(u+ct)dt. \tag{4.4}$$

Proof. By conditional on the time and the amount of the first claim, we determine that

$$G(u,b) = \int_0^{\frac{b-u}{c}} \int_0^{u+ct} G(u+ct-x,b)f_{X,W}(x,t)dxdt + \int_0^{\frac{b-u}{c}} \int_{u+ct}^\infty f_{X,W}(x,t)dxdt \tag{4.5}$$

exchanging (1.7) in (4.5) leads to

$$\begin{aligned} G(u,b) = &\int_0^{\frac{b-u}{c}} \int_0^{u+ct} G(u+ct-x,b)\lambda e^{-\lambda t}f(x)dxdt \\ &+ \theta \int_0^{\frac{b-u}{c}} \int_0^{u+ct} G(u+ct-x,b)g(t)h(x)dxdt \\ &+ \int_0^{\frac{b-u}{c}} \int_{u+ct}^\infty \lambda e^{-\lambda t}f(x)dxdt \\ &+ \theta \int_0^{\frac{b-u}{c}} \int_{u+ct}^\infty g(t)h(x)dxdt. \end{aligned}$$

By (4.2), (4.3) and (4.4) we have

$$G(u,b) = \int_0^{\frac{b-u}{c}} \lambda e^{-\lambda t}\gamma_1(u+ct)dt + \theta M(u).$$

With $u+ct = s$, we have

$$G(u,b) = \int_u^b \lambda e^{-\lambda\frac{s-u}{c}}\gamma_1(s)\frac{1}{c}ds + \theta M(u). \tag{4.6}$$

Differentiation the two sides of (4.6) w.r.t u, we acquire

$$\frac{\partial G(u,b)}{\partial u} = \int_u^b \frac{\lambda^2}{c}e^{-\lambda\frac{s-u}{c}}\gamma_1(s)\frac{1}{c}ds - \frac{\lambda}{c}\gamma_1(u) + \theta.\acute{M}(u), \tag{4.7}$$

thanks to (4.6) and (4.7) can be rewritten as

$$\left(\frac{\lambda}{c}\mathcal{I} - \mathcal{D}\right)G(u,b) = \frac{\lambda}{c}\theta M(u) + \frac{\lambda}{c}\gamma_1(u) - \theta\acute{M}(u). \tag{4.8}$$

Differentiation (4.8) w.r.t u once again and with some rearrangements, we obtain

$$\left(\frac{\lambda}{c}\mathcal{D}-\mathcal{D}^2\right)G(u,b)=\frac{\lambda}{c}\theta\acute{M}(u)+\frac{\lambda}{c}\acute{\gamma}_1(u)-\theta\acute{M}(u), \tag{4.9}$$

and with help (4.8) the equation (4.9) leads to (4.1). □

Remark 4.2. In particular, if $\phi_1(t)=\phi_2(t)=1-t$, (4.1) satisfies equation (3.1) in [5] which relates to the consequence of the distribution of the maximum surplus before destroy $G(u,b)$ when X and W are dependent with FGM copula function.

Remark 4.3. If $\theta=0$, (4.9) coincides with equation (2.6) when $n=1$ in [7] and (4.1) relates to the result of the distribution of the maximum surplus before failure $G(u,b)$ when X is independent of W as in the classical compound Poisson risk model.

Remark 4.4. Furthermore, by taking the Laplace transform of (4.1), we may comprehend this second order differential equation and determine a correct portrayal for $G(u,b)$, which is additionally the solution of a defective renewal equation.
But $G(u,b)$ can be assessed through this defective renewal function only for a few special choices of copula functions, and special choices of claim amount distributions combinations of exponential [5], a mixture of Erlangs, etc., since its representation is rather involved. Usually, we could only obtain asymptotic for $G(u,b)$ for general claim amount distributions.

Definition 1. *Suppose that for* $0\leq u\leq b$

$$\tau^b=\inf\{t>0,R(t)\geq b|R(0)=u\}, \tag{4.10}$$

to be the first time that the surplus process up crosses the level b, and

$$\chi(u,b)=P(T>\tau^b|R(0)=u)$$

to be the probability that the surplus process attains a given level b from initial surplus u without first falling below zero. Since eventually either failure occurs without the surplus process attaining level b or the surplus attains level b, then we have $\chi(u,b)=1-G(u,b)$.

Proposition 4.5. *For* $0\leq u\leq b$, *the probability* $\chi(u,b)$ *satisfies the Integro-differential equation*

$$\left(-\mathcal{D}^2+\frac{\lambda^2}{c^2}\mathcal{I}\right)\chi(u,b)=\left(\frac{\lambda^2}{c^2}\mathcal{I}+\frac{\lambda}{c}\mathcal{D}\right)\mu_1(u)-\theta\left(\frac{\lambda^2}{c^2}\mathcal{I}-\mathcal{D}^2\right)L(u), \tag{4.11}$$

where

$$\mu_1(u)=\int_0^u\chi(u-x,b)f(x)dx, \tag{4.12}$$

$$\mu_2(u) = \int_0^u \chi(u-x,b)h(x)dx, \tag{4.13}$$

and

$$L(u) = \int_0^{\frac{b-u}{c}} g(t)\mu_2(u+ct)dt. \tag{4.14}$$

5 Conclusions

In this paper, we have demonstrated that some techniques that are famous can be used to solve the probability of maximum surplus before failure $G(u,b)$ under classical compound Poisson risk model where the claim sizes depend on inter-claim times via a Sarmanov copula. We refer the reader to [8] for details.

References

[1] Albrecher, H., Teugels, J. (2006). *Exponential behavior in the presence of dependence in risk theory*. Journal of Applied Probability 43, 265285.

[2] Cheung, E. C. K.Landriault, D. (2009). *Perturbed MAPriskmodels with dividend barrier strategies*. Journal of Applied Probability 46, 521541.

[3] Cossette,H., Marceau, E.,Marri, F. (2008).*On the compound Poisson riskmodelwith dependence based on a generalized FarlieGumbelMorgenstern copula*. Insurance: Mathematics and Economics 43, 444455.

[4] Cossette, H., Marceau, E. & Marri, F. (2010). *Analysis of ruine measures for the classical compound Poisson risk model with dependence*. Scandinavian Actuarial Journal 2010, 221245.

[5] Jiang, W., Yang, Z. (2016). *The maximum surplus before ruine for dependent risk models through FarlieGumbelMorgenstern copula*. Scandinavian Actuarial Journal, 2016(5), 385-397.

[6] Jiang, W. Y., Yang, Z. J. & Li, X. P. (2012). *The discounted penalty function with multi-layer dividend strategy in the phase-type risk model*. Statistics and Probability Letters 82, 13581366.

[7] Li, S., Dickson, D. C. (2006). *The maximum surplus before ruine in an Erlang (n) risk process and related problems*. Insurance: Mathematics and Economics, 38(3), 529-539.

[8] Li, S., Lu, Y. (2010). *On the maximum severity of ruine in the compound Poisson model with a threshold dividend strategy*. Scandinavian Actuarial Journal, 2010(2), 136-147.

[9] Lin, X. S. & Sendova, K. P. (2008). *The compound Poisson risk model with multiple thresholds*. Insurance: Mathematics and Economics 42, 617627.

[10] McNeil, A. J., Frey, R., Embrechts, P. (2005). *Quantitative risk management*: Concepts, techniques and tools (Vol. 3). Princeton: Princeton university press.

[11] Nikoloulopoulos, A. K., Karlis, D. (2008). *Fitting copulas to bivariate earthquake data*: the seismic gap hypothesis revisited. Environmetrics: The official journal of the International Environmetrics Society, 19(3), 251-269.

[12] Zhang, Z. M. & Yang, H. (2011). *GerberShiu analysis in a perturbed risk model with dependence between claim sizes and interclaim times*. Journal of Computational and Applied Mathematics 235, 11891204.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Joint Modelling of Longitudinal and Survival Data Using Copulas

Kazempoor, J. [1] Habibirad, A. [2]

Department of Statistics, Ferdowsi University of Mashhad, Iran

Abstract

In this study, joint modeling of two longitudinal and survival sub models have been considered. We utilize two student-t and Gaussian copulas for joint modeling of these submodels. The parameter of these sub models under the main joint model is estimated using maximizing likelihood estimation method. Finally, from a simulation strategy, the performance of this model is indicated.

Keywords: Copula , Joint Modeling , Longitudinal , Submodel , Survival.

1 Introduction

Assume that a case which we are interested in assessing some of its behaviour during the random time or during the time until a special event has occurred. All of these data which were taken is longitudinal and the random specific event time is survival data. The proposed model for longitudinal data is often linear, linear mixed effect, exponential, logistic model

[1]kazempoor.jaber@mail.um.ac.ir
[2]ahabibi@um.ac.ir

and so forth (see [1],[2] and [5]). For survival data, some frailty models can be used such as gamma, Weibull and Cox frailty models (see [3] and [8]).

It is because each of these data observed in one individual, it is rational to consider some dependency between these data. Consequently, joint modelling of these sub models is so important from an aspect for better predicting time to the event of each case based on its longitudinal measurements. There are many strategies for joint modelling of longitudinal and survival data (see [1],[3],[2], [5] and [8]).

In the present study, the joint modeling of these data is constructed based on student-t copula and the model of longitudinal data has been considered arising from Gaussian copulas. The first choice can be changed according to the researcher's desire, but for longitudinal data, we must use the Gaussian copula because of these data considering to following from a multivariate normal distribution (see [1],[3] and[2]).

In the following, the process of constructing the main likelihood function of the joint modeling of longitudinal and survival submodels through student-t and Gaussian copulas are extensively explained in section 2. Section 3 has paid to the simulation study of the performance of Maximum Likelihood estimators (MLE) and finally, in section 4 we provide an applicable related example.

2 Theoretical Backgrounds

Suppose there are n subjects in a study, where the ith subject has n_i longitudinal measurements y_{ij} at time t_{ij}. We consider linear mixed effect model in the form $Y_i = X_iB+Z_i\nu_i+\epsilon_i$ as a longitudinal sub model where X_i and Z_i respectively represents the design matrix of fixed and random effects for each individuals. In addition, two vectors ν_i and ϵ_i assumed to be mutually independent and follows from $N(0,\sigma_\nu^2)$ and $N(0,\sigma_\epsilon^2)$ distributions. The random effects utilized as a showcase of the unknown effects of each individual on the repeated measurements that can not be uttered by the observed covariates. It is clear to see that $Y_i \sim N(X_iB, \sigma_\nu^2 Z_i Z_i' + \sigma_\epsilon^2)$

In the case of survival data for simplicity, one can used the two parameters Weibull frailty sub model as $h_{S_i}(s_i) = \frac{\alpha}{\lambda^\alpha} s_i^{\alpha-1}$. If the censored component has been considered any positive constant value c, the observed survival data is $T_i = max(S_i, c)$. In order to derivation probability density function (pdf) and cumulative density function (cdf) of T_i, if the cdf of random variable S_i is considered to following absolutely continuous cdf F, then with some slight mathematical calculation, we can get

$$\begin{aligned} F_{T_i}(t_i) &= F_{S_i}(t_i)I_{(c,\infty]}(t_i) + F_{S_i}(c)I_{[c]}(t_i), \\ f_{T_i}(t_i) &= f_{S_i}(t_i)I_{(c,\infty)}(t_i) + F_{S_i}(c)I_{[c]}(t_i). \end{aligned}$$

Since $F_{S_i}(t) = 1 - e^{-(\frac{t}{\lambda})^\alpha}$ and $f_{S_i}(t) = \frac{\alpha}{\lambda^\alpha} t^{\alpha-1} e^{-(\frac{t}{\lambda})^\alpha}$, we have

$$\begin{aligned} F_{T_i}(t_i) &= (1 - e^{-(\frac{t_i}{\lambda})^\alpha}) I_{(c,\infty]}(t_i) + 1 - e^{-(\frac{c}{\lambda})^\alpha} I_{[c]}(t_i), \\ f_{T_i}(t_i) &= \frac{\alpha}{\lambda^\alpha} {t_i}^{\alpha-1} e^{-(\frac{t_i}{\lambda})^\alpha} I_{(c,\infty)}(t_i) + (1 - e^{-(\frac{c}{\lambda})^\alpha}) I_{[c]}(t_i). \end{aligned}$$

Now, we consider joint model of observed data Y_i, T_i as a copula of marginal cumulative distribution function of survival data and copula of joint cdf of observed data in the form of $C_1(F_{S_i}(t_i), C_2(F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i})))$. The kind of C_2 copula is n_i dimensional Gaussian because of the relation $Y_i \sim N(X_i\beta, \sigma_\nu^2 ZZ' + \sigma_\epsilon^2 I_{n_i \times n_i})$ and for the C_1 the elliptical student-t copula has been considered.

Here, utilize the sklar's theorem ([4] and [6]) to get the joint cdf of Y_i, T_i as

$$\begin{aligned} C_1(F_{T_i}(t_i), F_{Y_{i1},\ldots,Y_{in_i}}(y_{i1}, \ldots, y_{in_i})) &= C_1(F_{T_i}(t_i), C_2(F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i}))) \\ &= C(F_{T_i}(t_i), F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i})) \end{aligned}$$

The existence of C as a copula can be easily checked in the dimension $n_i = 2$ and consequently proved for any dimensions inductively. From another point of view there exist a copula, C^* say, such that

$$F_{T_i,Y_{i1},\ldots,Y_{in_i}}(t_i, y_{i1}, \ldots, y_{in_i}) = C^*(F_{T_i}(t_i), F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i}))$$

The specification of C^* does not clearly seem. The choice of C_2 is fixed and the choice of C_1 can lead us to finding better alternatives for C^*. Here, we choose student-t copula as a supersede of C_1 and then aim to present how does it work through a simulation study.

The construction of two dimensional student-t copula density function with $k > 2$ degrees of freedom is as follow. Firstly suppose that two random variables U_1 and U_2 following a two dimensional student-t pdf with $k > 2$ degrees of freedom such that

$$f^k_{U_1,U_2}(u_1, u_2) = \frac{\sqrt{det(A)}}{2\pi} (1 + \frac{(u-\mu)' A (u-\mu)}{k})^{-(\frac{k}{2}+1)},$$

where

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \mu = \begin{bmatrix} E(U_1) \\ E(U_2) \end{bmatrix}, \begin{bmatrix} Var(U_1) & Cov(U_1, U_2) \\ Cov(U_1, U_2) & Var(U_2) \end{bmatrix} = \frac{k}{k-2} A^{-1}$$

$$A = \frac{8k}{k-2} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \mu = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Here, the two dimensional student-t copula density function with $k > 2$ degrees of freedom constructed based on this pdf has been considered.

$$c_1(u, w) = \frac{f^k_{U_1,U_2}(t_k^{-1}(u), t_k^{-1}(w))}{f^k_{U_1}(t_k^{-1}(u)) f^k_{U_2}(t_k^{-1}(w))},$$

where $f_U^k(.)$ represent the one dimensional student-t pdf with k degrees of freedom which defined as

$$f_U^k(u) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{k\pi}}(1+\frac{u^2}{k})^{-\frac{k+1}{2}},$$

and $t_k^{-1}(.)$ denotes the quantile function of a standard univariate student-t distribution with k degrees of freedom and satisfy in the relation $\int_{-\infty}^{t_k^{-1}(z)} f_U^k(u)du = z$. The joint probability density function of obsereved data in ith individual is

$$\begin{aligned} f_{T_i,Y_{i1},\ldots,Y_{in_i}}(t_i, y_{i1}, \ldots, y_{in_i}) &= c_1(F_{T_i}(t_i), C_2(F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i}))) \\ &\times \; c_2(F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i}))f_{T_i}(t_i)\prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij}) \\ &= c_1(F_{T_i}(t_i), C_2(F_{Y_{i1}}(y_{i1}), \ldots, F_{Y_{in_i}}(y_{in_i})))f_{T_i}(t_i) \\ &\times \; \frac{\sqrt{2\pi}\phi[(Y_i - X_iB)'\Sigma(Y_i - X_iB)]}{\sqrt{(2\pi)^{n_i}det(\Sigma)}} \end{aligned}$$

where ϕ and Φ respectively denotes the pdf and cdf of univariate standard normal distribution.

Finally, as the last part of this section, with the help of the independence of each of the samples, the maximum likelihood function of observation is constructed as follows

$$L(\alpha, \lambda, \underline{B}, \sigma_\nu^2, \sigma_\epsilon^2) = \prod_{i=1}^{n} f_{T_i,Y_{i1},\ldots,Y_{in_i}}(t_i, y_{i1}, \ldots, y_{in_i}).$$

3 Simulation Study

In order to study the numerical behavior of copulas joint modelling of survival and longitudinal sub models, we ran a series of simulations as follows:

Firstly, consider longitudinal data Y_{ij} collected from specific individual i according to the time t_{ij} and kind of treatment x_{ij} in that time untill an event has occured. For this submodel consider linear mixed models as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}x_{ij} + \nu_i + \epsilon_{ij}, \qquad i = 1, 2, \ldots, n, \;\; j = 1, 2, \ldots, n_i$$

where the random effects ν_i were simulated from $N(0, \sigma_\nu^2)$ and the random errors ϵ_{ij} were generated from $N(0, \sigma_\epsilon^2)$. t_{ij} denotes the follow up times of the ith individual were taken between baseline time 1 and survival time T_i. x_{ij} is also represent control group 0 and treatment group 1 which is generally considered in equal number of each case but their order

may be change in some application problem. It is straightforward to assess that longitudinal random follow a normal distribution or for simplicity $F_{Y_{ij}}(y_{ij}) = \Phi(\frac{y_{ij}-\mu_{ij}}{\sqrt{\Sigma_{ij}}})$, where

$$\mu_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij} x_{ij}, \Sigma_{n_i \times n_i} = \begin{bmatrix} \sigma_\nu^2 + \sigma_\epsilon^2 & \sigma_\nu^2 & \dots & \sigma_\nu^2 \\ \sigma_\nu^2 & \sigma_\nu^2 + \sigma_\epsilon^2 & \dots & \sigma_\nu^2 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \sigma_\nu^2 + \sigma_\epsilon^2 & \sigma_\nu^2 & \dots & \sigma_\nu^2 + \sigma_\epsilon^2 \end{bmatrix}$$

The survival time T_i is assumed to be $max(S_i, c)$ where S_i were generated from Weibull frailty model with shape parameter α and scale parameter λ. Finally, the replicate of these simulation study has been considered m.

Our algorithm of these simulation contains the following steps:

I: Fix longitudinal sub model parameters $\beta_0, \beta_1, \beta_2$, survival sub model parameters α, λ, c, number of individuals n, variance of errors σ_ϵ^2 and variance of random effects σ_ν^2

II: Generate n ,n and $\sum_{i=1}^{n} n_i$ random variables S_i ,ν_i and ϵ_{ij} from distributions $Weibull(\alpha, \lambda)$,$N(0, \sigma_\epsilon^2)$,$N(0, \sigma_\nu^2)$ respectively.

III: Construct x as a vector of size n contains equal number of 0 and 1 sequently.

IV: Construct $\sum_{i=1}^{n} n_i \times n$ matrix Z contain arrays $Z_{ji} = 1$, $i = 1, 2, \dots, n$, $n_{i-1} + 1 \leq i \leq n_i$ where $n_0 = 0$ and 0 otherwise.

V: Put $t_{ij} = 1, 2, \dots, [T_i]$ where $T_i = max(S_i, c)$ and notation [.] present greatest integer value less or equal to the containing number.

VI: Construct $\sum_{i=1}^{n} n_i \times 3$ matrix X contain 3 columns, the first is equal to $1_{\sum_{i=1}^{n} n_i \times 1}$, the second is equal to t_{ij} and the last is the form of $x_{ij} \times t_{ij}$.

VII: Constitute observed data matrix Y as $Y = XB + Z\nu + \epsilon$.

VIII: Consider lakilihood function as a function of $\alpha, \lambda, b0, b1, b2, \sigma_\nu^2, \sigma_\epsilon^2$ and calculate MLE of these parameters

IX: Repeat steps $\Big(\text{II}\Big)$-$\Big(\text{VIII}\Big)$ m times for deriving mean squared errors (MSE), mean absolute errors (MAE) and bias (Bias) of these estimations.

In continued, we provide some tables for a deep grasp of the performance of this joint modeling.

$c = 19,\ df = 5,\ n = 25,\ m = 1000$			
.	Bias	MSE	MAE
$\alpha = 3$	0.547	1.513	0.952
$\lambda = 26$	-0.637	1.907	1.193
$\beta_0 = 0.9$	-0.128	0.933	0.769
$\beta_1 = 0.05$	0.098	0.370	0.268
$\beta_2 = 0.8$	-0.176	0.441	0.357
$\sigma^2_\nu = 0.05$	0.210	0.247	0.163
$\sigma^2_\epsilon = 0.01$	-0.111	0.912	0.640

$c = 19,\ df = 5,\ n = 50,\ m = 500$			
.	Bias	MSE	MAE
$\alpha = 3$	0.563	1.650	0.976
$\lambda = 26$	-0.788	1.877	1.088
$\beta_0 = 0.9$	-0.169	0.969	0.898
$\beta_1 = 0.05$	0.067	0.299	0.257
$\beta_2 = 0.8$	-0.163	0.416	0.330
$\sigma^2_\nu = 0.05$	0.180	0.137	0.120
$\sigma^2_\epsilon = 0.01$	-0.150	0.842	0.577

$c = 19,\ df = 5,\ n = 100,\ m = 250$			
.	Bias	MSE	MAE
$\alpha = 3$	0.520	1.465	0.973
$\lambda = 26$	-0.624	1.682	0.990
$\beta_0 = 0.9$	-0.154	0.993	0.748
$\beta_1 = 0.05$	0.076	0.274	0.219
$\beta_2 = 0.8$	-0.141	0.359	0.317
$\sigma^2_\nu = 0.05$	0.147	0.116	0.109
$\sigma^2_\epsilon = 0.01$	-0.121	0.794	0.548

It is straightforward to see that even we do not use shared frailty models for each case, the performance of estimators based on this joint model not bad at all. In addition, it can be mentioned that these results for $n = 100,\ m = 500$ and $n = 100,\ m = 1000$ is much better than the results given here.

The main point is that for degrees of freedom less than 10, $n > 50$ and $m > 100$ the results look good.

4 Applications

In this section, we aim to predict champion of half season $(2017 - 2018)$ of England premier league. The survival time is that at least one team reach 57 points and must be at least 19

games. Control and treatment group in this situation interpreted as each team gaming in home 1 or away 0. According to the history of premier league all of top six teams have been considered and based on their previous season results, can be immediately understood that the survival time of Manchester United, Arsenal, Chelsea, Manchester City, Liverpool, and Tottenham Hotspur is 28, 34, 33, 20, 28 and 29 respectively. The number of scored goals and place of each game are provided in the following table.

2017 – 2018						
.	Arsenal	Manchester United	Manchester City	Liverpool	Tottenham Hotspur	Chelsea
1	1-4	1-4	0-2	0-3	0-2	1-2
2	0-0	0-4	1-1	1-1	1-1	0-2
3	0-0	1-2	0-2	1-4	1-1	1-2
4	1-3	0-2	1-5	0-0	0-3	0-2
5	0-0	1-4	0-6	1-1	1-0	1-0
6	1-2	0-1	1-5	0-3	0-3	0-4
7	1-2	1-4	0-1	0-1	0-4	1-0
8	0-1	0-0	1-7	1-0	1-1	0-1
9	0-5	0-1	1-3	0-1	1-4	1-4
10	1-2	1-1	0-3	1-3	0-0	0-1
11	0-1	0-0	1-3	0-4	1-1	1-1
12	1-2	1-4	0-2	1-3	0-0	0-4
13	0-1	1-1	0-2	1-1	1-1	0-1
14	1-5	0-4	1-2	0-3	0-1	1-1
15	1-1	0-3	1-2	0-5	0-1	1-3
16	0-1	1-1	0-2	1-1	1-5	0-0
17	0-0	1-0	0-4	1-0	1-2	0-3
18	1-1	0-2	1-4	0-4	0-1	1-1
19	1-3	0-2	1-4	0-3	0-3	0-0
20	0-3	1-2	0-1	1-5	1-5	1-2
21	0-1	1-0	-	1-2	0-2	1-5
22	1-2	0-2	-	0-2	1-1	0-2
23	0-1	1-3	-	1-4	1-4	1-0
24	1-4	0-1	-	0-0	0-1	0-4
25	0-1	0-0	-	0-3	1-2	1-0
26	1-5	1-2	-	1-2	0-2	0-1
27	0-0	0-0	-	0-2	1-1	1-3
28	1-0	1-2	-	1-4	0-1	0-1
29	0-1	-	-	-	1-2	0-0
30	1-3	-	-	-	-	1-2
31	1-3	-	-	-	-	1-1
32	1-3	-	-	-	-	1-1
33	0-1	-	-	-	-	0-3
34	1-4	-	-	-	-	-

According to these data, the MLE of our parameters is $\hat{\alpha} = 6.896,\ \hat{\lambda} = 29.321,\ \hat{b}_0 = 1.041,\ \hat{b}_1 = 0.048,\ \hat{b}_2 = 0.079,\ \hat{\sigma}_\nu = 0.141, \hat{\sigma}_\epsilon = 0.052$. Here with respect to these estimators, the number of scored goals of these six teams considering the position of the start game, until reaching 57 points in the current season are predicted as follow

2018 – 2019						
.	Arsenal (Home)	Manchester United (Home)	Manchester City (Away)	Liverpool (Home)	Tottenham Hotspur (Away)	Chelsea (Away)
1	1	2	2	3	2	1
2	2	2	2	2	2	3
3	3	1	1	1	3	2
4	2	1	2	2	1	2
5	2	2	3	1	0	3
6	3	1	1	3	2	0
7	2	2	2	2	1	2
8	1	3	0	0	1	3
9	3	2	3	1	2	2
10	2	2	1	3	1	3
11	2	1	2	2	3	1
12	1	1	3	2	2	0
13	1	0	1	3	3	1
14	4	0	3	2	2	1
15	2	2	2	3	3	1
16	2	4	0	1	3	2
17	1	1	2	2	1	2
18	0	2	4	4	1	3
19	1	2	3	2	3	2
20	2	3	2	1	2	1
21	3	0	2	2	1	0
22	1	1	1	-	0	2
23	0	1	0	-	2	2
24	4	2	3	-	1	3
25	2	2	-	-	0	1
26	3	0	-	-	2	0
27	0	1	-	-	3	1
28	1	4	-	-	1	2
29	1	3	-	-	-	1
30	3	1	-	-	-	-
31	-	1	-	-	-	-
32	-	1	-	-	-	-
33	-	2	-	-	-	-
34	-	3	-	-	-	-

References

[1] Diggle, P. Diggle, P. J. Heagerty, P. Heagerty, P. J. Liang, K. Y. and Zeger, S. (2002). *Analysis of longitudinal data*, Oxford University Press.

[2] Fitzmaurice, G. Davidian, M. Verbeke, G. and Molenberghs, G. (2008), *Longitudinal data analysis*, CRC press.

[3] Hougaard, Philip. (1995), Frailty models for survival data, *Lifetime data analysis* **1** 255-273.

[4] Joe, H. (2004), *Dependence modeling with copulas*, Chapman and Hall/CRC.

[5] Liang, K. Y. and Scott, L. Z. (1986), Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13-22.

[6] Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

[7] Sattar, A. and Sinha, S.K. (2017), Joint modelling of longitudinal and survival data with a covariate subject to a limit of detection, *Statistical methods in medical research*, DOI: 10.1177/0962280217729573.

[8] Wienke, A. (2010), *Frailty models in survival analysis*, Chapman and Hall/CRC.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Discrete Bivariate Distributions Generated By Copulas: DBEEW Distribution

Najarzadegan, H. [1] Alamatsaz, M.H. [2] Kazemi, I. [3]

Department of Statistics, University of Isfahan, Isfahan, Iran

Abstract

In this paper, we shall propose a general method of generating discrete bivariate distributions using copulas. The advantage of our method is that, contrary to the standard methods, we do not need to have the joint distribution of the base variables, we need the marginals only. In particular, we shall concentrate on generating a new discrete bivariate exponentiated extended Weibull (DBEEW) by a Cuardas-Auge copula. An advantage of this family of copulas is that they are exchangeable and, thus, can cover exchangeable random vectors which are widely used in Engineering fields.

Keywords: Cuardas-Auge copulas, Discrete bivariate distribution, Probability generating function.

[1]hnajarzadegan@yahoo.com
[2]alamatho@sci.ui.ac.ir
[3]ikazemi@sci.ui.ac.ir

1 Introduction

Due to the limited number of flexible discrete distributions, in recent years, several methods have been proposed to construct new discrete distributions. One of the most important methods is the discretization of known continuous distributions. Let $h(y_1, y_2)$ and $H(y_1, y_2)$ be the joint pdf and cdf of a given bivariate contiuous distribution on $(0, \infty)^2$. Then, we can generate a new discrete bivariate distribution by the following 2 methods

$$P(Y_1 = y_1, Y_2 = y_2) = \frac{h(y_1, y_2)}{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h(i, j)}, \qquad y_1, y_2 = 1, 2, ..., \tag{1.1}$$

or

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2) =& H(y_1 + 1, y_2 + 1) - H(y_1 + 1, y_2) - H(y_1, y_2 + 1) \\ &+ H(y_1, y_2), \qquad y_1, y_2 = 0, 1, ..., \end{aligned} \tag{1.2}$$

Clearly, relations (1.1) and (1.2) provide new discrete bivariate distributions provided that the joint continuous distribution of Y_1 and Y_2 is known. Motivated by (1.2), we can develope a new method of generating discrete bivariate distribution using copula functions having only the maginal continuous distributions of Y_1 and Y_2.
To consider the type of the dependence structure between the variables, copulas play important roles in probability theory and statistics. Recall that copulas are probability functions which connect multivariate distributions to their marginal distributions. Sklar [7] proved that there is a unique copula function C connecting continious marginal distributions G_1 and G_2 to their joint continuous bivariate distributions H as

$$C(G_1(y_1), G_2(y_2)) = H(y_1, y_2). \tag{1.3}$$

Now, inserting (1.3) into (1.2) we obtain new family of discrete bivariate distributions generated by copulas as

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2) =& C\big(G_{Y_1}(y_1 + 1), G_{Y_2}(y_2 + 1)\big) - C\big(G_{Y_1}(y_1 + 1), G_{Y_2}(y_2)\big), \\ &- C\big(G_{Y_1}(y_1), G_{Y_2}(y_2 + 1)\big) + C\big(G_{Y_1}(y_1), G_{Y_2}(y_2)\big). \end{aligned} \tag{1.4}$$

It is obvious that a great advantage of this method is that we need not have the parent joint distribution and given two marginals we can produce a large class of bivariate discrete distributions by usin different copulas C in (1.4). Some known parametric copulas are the Farlie-Gumbel-Morgenstern (FGM) copulas, Normal copulas, the Marshall and Olkin copulas, Cuadras-Auge copulas, etc.
The Cuardas-Auge family of copulas was introduced based on the family of bivariate distributions proposed by Cuadras and Auge [2] as

$$C_\theta(u, v) = \min\{u, v\}. \max\{u, v\}^{1-\theta} \tag{1.5}$$

where $0 \leq \theta \leq 1$ and $0 \leq u, v \leq 1$. The parameter θ measures the degree of dependence, where $\theta = 0$ corresponds to independence and $\theta = 1$ to complete comonotonicity. Moreover, θ plays the role of the parameter of upper-tail dependence of $C_\theta(u, v)$. The Cuardas-Auge copulas is exchangeable, i.e. symmetric in its marginals. For this reason, the Cuardas-Auge copulas are largely used in a variety of modelings for exchangeable random vectors. Thus, considering the dependency structure between Y_1 and Y_2, we can choose the Cuardas-Auge copula in (1.4) to produce our desired joint distribution. By inserting the Cuardas-Auge copulas (1.5) into Eq. (1.4) for any marginal univariate continious cdf $G_1(.) = G_2(.) = G(.)$ a new large family of discrete bivariate distribution can be drived as

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \Big[\big(G(y_1+1)\big)^\gamma - \big(G(y_1)\big)^\gamma\Big]\Big[G(y_2+1) - G(y_2)\Big], & \text{if } y_1 > y_2, \\ \Big[\big(G(y_2+1)\big)^\gamma - \big(G(y_2)\big)^\gamma\Big]\Big[G(y_1+1) - G(y_1)\Big], & \text{if } y_2 > y_1, \\ \big(G(y+1)\big)^\gamma\Big[G(y+1) - G(y)\Big] - G(y)\Big[\big(G(y+1)\big)^\gamma - \big(G(y)\big)^\gamma\Big] & \text{if } y_1 = y_2 = y \end{cases} \tag{1.6}$$

where $y_1, y_2 = 0, 1, ...$ and $\gamma = 1 - \theta$.
The exponentiated extended Weibull (EEW) distribution is a popular continuous family of distributions whose cumulative distribution function is given by

$$G_{EEW}(x; \alpha, \lambda, \boldsymbol{\xi}) = \big(1 - e^{-\lambda H(x;\boldsymbol{\xi})}\big)^\alpha, \quad x > 0, \alpha > 0, \lambda > 0, \tag{1.7}$$

where $H(x; \boldsymbol{\xi})$ is a non-negative increasing function depending on parameter vector $\boldsymbol{\xi} > 0$. Most bivariate distributions with EEW marginals are all defined on continuous scales. But, in real cases, we usually encounter situations which can not measure the life length of a device on a continuous scales. Therefore, in this paper we intend to construct new discrete bivariate distributions by using EEW distribution. Our main aim now is to propose a new discrete bivariate exponentiated extended Weibull (DBEEW) distribution by using EEW distribution as the indentical parent cdf G in Eq. (1.6).

2 The DBEEW Distribution And Its Basic Properties

Now by utilizing the exponentiated extended Weibull (EEW) cdf (1.7) in Eq. (1.6) we shall arrive at a new discrete bivariate exponentiated extended Weibull (DBEEW) distribution with pmf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} P_1(y_1, y_2), & \text{if } y_1 > y_2, \\ P_2(y_1, y_2), & \text{if } y_2 > y_1, \\ P_0(y_1, y_2) & \text{if } y_1 = y_2 = y \end{cases} \tag{2.1}$$

where

$$
\begin{aligned}
P_1(y_1,y_2) =& \Big[\big(1-p^{H(y_1+1;\boldsymbol{\xi})}\big)^{\alpha\gamma} - \big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big]\Big[\big(1-p^{H(y_2+1;\boldsymbol{\xi})}\big)^{\alpha} - \big(1-p^{H(y_2;\boldsymbol{\xi})}\big)^{\alpha}\Big],\\
P_2(y_1,y_2) =& \Big[\big(1-p^{H(y_2+1;\boldsymbol{\xi})}\big)^{\alpha\gamma} - \big(1-p^{H(y_2;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big]\Big[\big(1-p^{H(y_1+1;\boldsymbol{\xi})}\big)^{\alpha} - \big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha}\Big],\\
P_0(y_1,y_2) =& \big(1-p^{H(y+1;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big[\big(1-p^{H(y+1;\boldsymbol{\xi})}\big)^{\alpha} - \big(1-p^{H(y;\boldsymbol{\xi})}\big)^{\alpha}\Big]\\
&- \big(1-p^{H(y;\boldsymbol{\xi})}\big)^{\alpha}\Big[\big(1-p^{H(y+1;\boldsymbol{\xi})}\big)^{\alpha\gamma} - \big(1-p^{H(y;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big],
\end{aligned}
$$

$y_1, y_2 \in N_0 = \{0,1,...\}$, $0 < p = e^{-\lambda} < 1$, $0 \le \gamma \le 1$ and $\alpha > 0$. We denote such a distribution by $DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$. It is observed that the corresponding joint cdf of a $(Y_1, Y_2) \sim DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$ is given by

$$
F_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha\gamma}\big(1-p^{H(y_2;\boldsymbol{\xi})}\big)^{\alpha\gamma}\big(1-p^{H(min\{y_1,y_2\};\boldsymbol{\xi})}\big)^{\alpha(1-\gamma)}, & y_1, y_2 = 0,1,..., \\ 0, & o.w. \end{cases} \tag{2.2}
$$

Survival function of a $DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$ random vector (Y_1, Y_2) can be obtained from the following equation

$$
\bar{F}_{Y_1,Y_2}(y_1,y_2) = 1 - F_{Y_1}(y_1) - F_{Y_2}(y_2) + F_{Y_1,Y_2}(y_1,y_2). \tag{2.3}
$$

Also, using the pmf (2.1) and survival function (2.3), one can obtain the bivariate hazard rate function as

$$
r(y_1,y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{\bar{F}_{Y_1,Y_2}(y_1,y_2)} \tag{2.4}
$$

where $y_1, y_2 \in N_0 = \{0,1,2,...\}$. Now we obtain the joint probability generating function (pgf) of a $DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$ distribution as follows

$$
\begin{aligned}
G_{Y_1,Y_2}(z_1,z_2) &= E(z_1^{Y_1} z_2^{Y_2}) = \sum_{y_2=0}^{\infty}\sum_{y_1=0}^{\infty} z_1^{y_1} z_2^{y_2} P(Y_1 = y_1, Y_2 = y_2),\\
&= \sum_{y_2=0}^{\infty}\sum_{y_1=y_2+1}^{\infty}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} z_1^{y_1} z_2^{y_2} (-1)^{i+j}\binom{\alpha\gamma}{i}\binom{\alpha}{j}(p^{iH(y_1+1;\boldsymbol{\xi})} - p^{iH(y_1;\xi)})(p^{jH(y_2+1;\boldsymbol{\xi})} - p^{jH(y_2;\boldsymbol{\xi})}),\\
&+ \sum_{y_1=0}^{\infty}\sum_{y_2=y_1+1}^{\infty}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} z_1^{y_1} z_2^{y_2} (-1)^{i+j}\binom{\alpha}{i}\binom{\alpha\gamma}{j}(p^{iH(y_1+1;\boldsymbol{\xi})} - p^{iH(y_1;\xi)})(p^{jH(y_2+1;\xi)} - p^{jH(y_2;\boldsymbol{\xi})}),\\
&+ \sum_{y=0}^{\infty}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} z_1^{y} z_2^{y} (-1)^{i+j}\binom{\alpha\gamma}{i}\binom{\alpha}{j} p^{iH(y+1;\boldsymbol{\xi})}(p^{jH(y+1;\boldsymbol{\xi})} - p^{jH(y;\boldsymbol{\xi})}),\\
&- \sum_{y=0}^{\infty}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} z_1^{y} z_2^{y} (-1)^{i+j}\binom{\alpha}{i}\binom{\alpha\gamma}{j} p^{iH(y;\boldsymbol{\xi})}(p^{jH(y+1;\boldsymbol{\xi})} - p^{jH(y;\boldsymbol{\xi})}),
\end{aligned} \tag{2.5}
$$

where $|z_1| < 1$ and $|z_2| < 1$. The marginal distributions of Y_1 and Y_2 can also be determined from the joint pgf (2.5) at $z_1 = 1$ and $z_2 = 1$, respectively.

Additionally, in the following theorem we shall obtain the conditional pmf, cdf and conditional expected value of the distribution without proof.

Theorem 2.1. *Let $(Y_1, Y_2) \sim DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$, then for all non-negative integer values of y_1 and y_2, we have*

(a) the conditional pmf of $(Y_2|Y_1)$,

$$f_{Y_2|Y_1=y_1}(y_2) = \begin{cases} \frac{P_1(y_1,y_2)}{(1-p^{H(y_1+1,\boldsymbol{\xi})})^\alpha-(1-p^{H(y_1,\boldsymbol{\xi})})^\alpha}, & \text{if } y_1 > y_2, \\ \frac{P_2(y_1,y_2)}{(1-p^{H(y_1+1,\boldsymbol{\xi})})^\alpha-(1-p^{H(y_1,\boldsymbol{\xi})})^\alpha}, & \text{if } y_2 > y_1, \\ \frac{P_0(y_1,y_2)}{(1-p^{H(y+1,\boldsymbol{\xi})})^\alpha-(1-p^{H(y,\boldsymbol{\xi})})^\alpha} & \text{if } y_1 = y_2 = y \end{cases} \tag{2.6}$$

(b) the conditional cdf of $(Y_2|Y_1 \leq y_1)$ is given by

$$F_{Y_2|Y_1\leq y_1}(y_2) = \frac{P(Y_2 \leq y_2, Y_1 \leq y_1)}{P(Y_1 \leq y_1)} = \begin{cases} (1-p^{H(y_1,\boldsymbol{\xi})})^{-\alpha(1-\gamma)}(1-p^{H(y_2,\boldsymbol{\xi})})^\alpha, & \text{if } y_1 > y_2, \\ (1-p^{H(y_2,\boldsymbol{\xi})})^{\alpha\gamma}, & \text{if } y_2 > y_1, \\ (1-p^{H(y,\boldsymbol{\xi})})^{\alpha\gamma} & \text{if } y_1 = y_2 = y \end{cases} \tag{2.7}$$

(c) the conditional cdf of $(Y_2|Y_1 = y_1)$ is given by

$$\begin{aligned} F_{Y_2|Y_1=y_1}(y_2) =& \frac{P(Y_1 = y_1, Y_2 \leq y_2)}{P(Y_1 = y_1)} \\ =& \begin{cases} \frac{\left(1-p^{H(y_2+1;\boldsymbol{\xi})}\right)^\alpha\left[\left(1-p^{H(y_1+1;\boldsymbol{\xi})}\right)^{\alpha\gamma}-\left(1-p^{H(y_1;\boldsymbol{\xi})}\right)^{\alpha\gamma}\right]}{\left[\left(1-p^{H(y_1+1;\boldsymbol{\xi})}\right)^\alpha-\left(1-p^{H(y_1;\boldsymbol{\xi})}\right)^\alpha\right]}, & \text{if } y_1 > y_2, \\ (1-p^{H(y_2+1,\boldsymbol{\xi})})^{\alpha\gamma}, & \text{if } y_2 > y_1, \\ \frac{\left(1-p^{H(y+1;\boldsymbol{\xi})}\right)^{\alpha\gamma}\left[\left(1-p^{H(y+1;\boldsymbol{\xi})}\right)^\alpha-\left(1-p^{H(y;\boldsymbol{\xi})}\right)^\alpha\right]}{\left[\left(1-p^{H(y_1+1;\boldsymbol{\xi})}\right)^\alpha-\left(1-p^{H(y_1;\boldsymbol{\xi})}\right)^\alpha\right]} & \text{if } y_1 = y_2 = y \end{cases} \end{aligned} \tag{2.8}$$

(d) the conditional expectation of $(Y_2|Y_1 = y_1)$ *is given by*

$$E(Y_2|Y_1=y_1) = \frac{\sum_{j=0}^{\infty}\sum_{y2=1}^{y_1-1} y_2\binom{\alpha}{j}(-1)^j\Big[\big(1-p^{H(y_1+1;\boldsymbol{\xi})}\big)^{\alpha\gamma}-\big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big]\Big[p^{jH(y_2+1;\boldsymbol{\xi})}-p^{jH(y_2;\boldsymbol{\xi})}\Big]}{(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha}}$$
$$-\frac{y_1\big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha}\Big[\big(1-p^{H(y_1+1;\boldsymbol{\xi})}\big)^{\alpha\gamma}-\big(1-p^{H(y_1;\boldsymbol{\xi})}\big)^{\alpha\gamma}\Big]}{(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha}}$$
$$+\sum_{j=0}^{\infty}\sum_{y2=y_1+1}^{\infty} y_2\binom{\alpha\gamma}{j}(-1)^j\Big[p^{jH(y_2+1;\boldsymbol{\xi})}-p^{jH(y_2;\boldsymbol{\xi})}\Big]+y_1\big(1-p^{H(y_1+1;\boldsymbol{\xi})}\big)^{\alpha\gamma} \tag{2.9}$$

Positively quadrant dependent and left-tail decreasing are two important properties in reliability topics. Random variables U_1 and U_2 are positively quadrant dependent if

$$F_{U_1,U_2}(u_1,u_2) \geq F_{U_1}(u_1)F_{U_2}(u_2) \tag{2.10}$$

for all u_1 and u_2 and U_2 is left- tail decreasing in U_1, if and only if, $F_{U_2|U_1\leq u_1}(u_2)$ is a non-increasing function of u_1 for any u_2. Therefore, based on these two definitions we have the following theorems.

Theorem 2.2. *Suppose* $(Y_1,Y_2) \sim DBEEW(\alpha,p,\gamma,\boldsymbol{\xi})$. *Then,* Y_1 *and* Y_2 *are positive quadrant dependen.*

Remark 2.3. Since Y_1 and Y_2 are positive quadrant dependen, then for any increasing functions $H_1(.)$ and $H_2(.)$ we have $Cov(H_1(Y_1),H_2(Y_2)) \geq 0$, (see [5]).

Theorem 2.4. *Suppose* $(Y_1,Y_2) \sim DBEEW(\alpha,p,\gamma,\boldsymbol{\xi})$. *Then,* Y_2 *is left- tail decreasing in* Y_1.

Some other more relibility measurs are the conditional hazard rate function, the hazard gradient vector and the Clayton-Oakes measure which we shall consider for BBEEW distribution as folllows. The conditional hazard rate function of $(Y_2|Y_1 = y_1)$ is given as the following theorem.

Theorem 2.5. *Let* $(Y_1,Y_2) \sim DBEEW(\alpha,p,\gamma,\boldsymbol{\xi})$. *Then*
(a) for $|y_1-y_2| \geq 1$ *we have*

$$r_{Y_2|Y_1}(y_2|y_1) = \begin{cases} r_1(y_2|y_1), & \text{if } y_1-y_2 \geq 1, \\ \\ r_2(y_2|y_1), & \text{if } y_1-y_2 \leq -1, \end{cases} \tag{2.11}$$

where

$$r_1(y_2|y_1) = \frac{P_1(y_1,y_2)}{[(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha}]-(1-p^{H(y_2+1,\boldsymbol{\xi})})^{\alpha}[(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha\gamma}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha\gamma}]}$$

and

$$r_2(y_2|y_1) = \frac{P_2(y_1,y_2)}{[1-(1-p^{H(y_2+1,\boldsymbol{\xi})})^{\alpha\gamma}][(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha}]}$$

(b) for $|y_1-y_2|<1$ *we have*

$$\begin{aligned} r_{Y_2|Y_1}(y_2|y_1) &= \frac{P_0(y_1,y_2)}{\sum_{i=y_1+1}^{\infty} P_2(Y_1=y_1,Y_2=i)} \\ &= \frac{P_0(y_1,y_1)}{[1-(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha\gamma}][(1-p^{H(y_1+1,\boldsymbol{\xi})})^{\alpha}-(1-p^{H(y_1,\boldsymbol{\xi})})^{\alpha}]} \end{aligned} \tag{2.12}$$

Similarly, we can obtain the conditional hazard rate function of $(Y_1|Y_2=y_2)$ which is useful to constract the hazard gradient which was proposed by Johnson and Kots (1975). They defined the hazard gradient as the vector

$$R(y_1,y_2) = (r_1(y_1,y_2), r_2(y_1,y_2))^T \tag{2.13}$$

where $r_1(y_1,y_1)$ is the hazard rate of the conditional distribution of Y_1 given $(Y_2>y_2)$ and $r_2(y_1,y_1)$ is the hazard rate of the conditional distribution of Y_2 given $(Y_1>y_1)$. Thus, using the conditional hazard rate functions and survival function (2.3) we can define the members of the hazard gradient of DBEEW distribution as

$$r_1(y_1,y_2) = r_{Y_1|Y_2>y_2}(y_1) = \frac{P(Y_1=y_1,Y_2>y_2)}{\bar{F}_{Y_1,Y_2}(y_1,y_2)} = \frac{r(y_1,y_2)}{r_{Y_2|Y_1}(y_2|y_1)} \tag{2.14}$$

and

$$r_2(y_1,y_2) = r_{Y_2|Y_1>y_1}(y_2) = \frac{P(Y_1>y_1,Y_2=y_2)}{\bar{F}_{Y_1,Y_2}(y_1,y_2)} = \frac{r(y_1,y_2)}{r_{Y_1|Y_2}(y_1|y_2)}, \tag{2.15}$$

respectively.
An association measure in reliability topics which was proposed by Oakes [6] and expanded by Clayton [1], is the Clayton−Oakes measure. This measure for DBEEW distribution is obtained by

$$\theta(x,y) = \frac{r_{Y_2|Y_1}(y_2|y_1)}{r_1(y_1,y_2)} \tag{2.16}$$

which can be simplified using Eq. (2.14) as

$$\theta(x,y) = \frac{r^2_{Y_2|Y_1}(y_2|y_1)}{r(y_1,y_2)}. \tag{2.17}$$

Now, here we consider several other important properties of $DBEEW(\alpha,p,\gamma,\xi)$ family in the following theorems.

Theorem 2.6. *Let* $(Y_1,Y_2)\sim DBEEW(\alpha,p,\gamma,\boldsymbol{\xi})$. *Then*

(a) $Y_{(1)} = \min\{Y_1, Y_2\}$ *has the cumulative distribution function*

$$F_{Y_{(1)}}(y) = \left(1 - p^{H([y];\boldsymbol{\xi})}\right)^{\alpha}\left(2 - \left(1 - p^{H([y];\boldsymbol{\xi})}\right)^{\alpha\gamma}\right), \tag{2.18}$$

(b) $Y_{(2)} = \max(Y_1, Y_2)$ *has the cumulative distribution function*

$$F_{Y_{(2)}}(y) = \left(1 - p^{H([y];\boldsymbol{\xi})}\right)^{\alpha(1+\gamma)}, \tag{2.19}$$

(c) $Y_2 - Y_1$ *has the probability mass function*

$$f_{Y_2-Y_1}(t) = \begin{cases} \sum_{y_1=0}^{\infty} P_2(y_1, y_1 + t), & t > 0, \\ \sum_{y_1=0}^{\infty} P_1(y_1, y_1 + t), & t < 0, \\ \sum_{y_1=0}^{\infty} P_0(y_1, y_1), & t = 0 \end{cases} \tag{2.20}$$

(d) $\left|Y_2 - Y_1\right|$ *has the probability mass function*

$$f_{\left|Y_2-Y_1\right|}(t) = \begin{cases} \sum_{y_1=0}^{\infty} P_2(y_1, y_1 + t) + \sum_{y_1=1}^{\infty} P_1(y_1, y_1 - t), & t > 0, \\ \sum_{y_1=0}^{\infty} P_0(y_1, y_1), & t = 0 \end{cases} \tag{2.21}$$

Theorem 2.7. *Let* $(Y_{j1}, Y_{j2}) \sim DBEEW(\alpha_j, p, \gamma, \boldsymbol{\xi})$*, for* $j = 0, 1, 2, ..., n$*, be independent random vectors and define* $X_1 = max(Y_{11}, Y_{21}, ..., Y_{n1})$ *and* $X_2 = max(Y_{12}, Y_{22}, ..., Y_{n2})$*. Then,* $(X_1, X_2) \sim DBEEW(\sum_{j=1}^{n} \alpha_j, p, \gamma, \boldsymbol{\xi})$

Theorem 2.8. *The stress-strenght probability of our proposed model* $(Y_1, Y_2) \sim DBEEW(\alpha, p, \gamma, \boldsymbol{\xi})$ *is given by*

$$P(Y_1 < Y_2) = \sum_{i=0}^{\infty} \left(1 - p^{H(i+1;\boldsymbol{\xi})}\right)^{\alpha}\left[\left(1 - p^{H(i+2;\boldsymbol{\xi})}\right)^{\alpha\gamma} - \left(1 - p^{H(i+1;\boldsymbol{\xi})}\right)^{\alpha\gamma}\right], \tag{2.22}$$

A discrete bivariate random vector (U_1, U_2) with pdf $f(\cdot, \cdot)$ is said to have a TP_2 property if

$$\frac{f(u_{11}, u_{21}) f(u_{12}, u_{22})}{f(u_{12}, u_{21}) f(u_{11}, u_{22})} \geq 1 \tag{2.23}$$

for all $u_{11} < u_{12}$ and $u_{21} < u_{22}$. Here we show that $DBEEW$ distribution does not have TP_2 property. To see this, consider a discrete bivariate generalized exponential (BDGE) distribution which is a special case of $DBEEW$ distribution. Then, choosing $x_{11} = 2 < x_{21} = 4 < x_{12} = 6 < x_{22} = 8$ and also $\alpha_1 = 5$, $\alpha_2 = 1$, $\alpha_3 = 3$ and $p = 0.9$, (24) yields

$$\frac{f(x_{11}, x_{21}) f(x_{12}, x_{22})}{f(x_{12}, x_{21}) f(x_{11}, x_{22})} = 0.9143 < 1.$$

Thus, BDGE and, consequently, the class of DBEEW distributions do not satisfy the TP_2 property in general.

Remark 2.9. The above observation shows that Nekoukhou and Kundu's [4] claim that BDGE distributions are TP_2 is not correct.

3 Some special cases

Three interesting sub-models of DBEEW class of distributions are discrete bivariate generalized exponential (DBGE) distribution ($H(y;\xi) = y$), discrete bivariate exponentiated Weibull (DBEW) distribution ($H(x;\xi) = x^{\beta}$) and discrete bivariate generalized Gompertz (DBGG) distribution ($H(x;\xi) = \beta^{-1}(e^{\beta x} - 1)$).

4 Application

We fit our three submodels of the DBEEW distribution and bivariate Poisson (BP) distribution to a real data set corresponding to Italian Series A football match score data between ACF Firontina and Juventus during 1996 to 2011, and conclude that all three and four parameters submodels provide better fits compared to bivariate Poisson distribution.

References

[1] Clayton, D.G. (1978), A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141151.

[2] Cuadras, C.M. and Auge, J. (1981), A continuous general multivariate distribution and its properties, *Communications in Statistics-Theory and Methods*, **10**(4), 339-353.

[3] Johnson, N.L. and Kotz, S. (1975), A vector valued multivariate hazard rate, *Journal of Multivariate Analysis*, **5**, 5366.

[4] Nekoukhou, V. and Kundu, D. (2017), Bivariate discrete generalized exponential distribution, *Statistics*, **51**(5), 1143-1158.

[5] Nelsen, R.B. (2006), *An introduction to copulas*, Springer, New York.

[6] Oakes, D. (1989), Bivariate survival models induced by frailties, *Journal of the American Statistical Association*, **84**, 487493.

[7] Sklar, M. (1959), Fonctions de repartition an dimensions et leurs marges, *Publications de lInstitut de statistique de lUniversite de Paris*, **8**, 229-231.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Joint Modeling of Correlated Responses in Insurance Data based on Gaussian Copula

Rezaee, F. [1] Bahrami Samani, E. [2] Ganjali, M. [3]

Department of Statistics, Shahid Beheshti University, Tehran, Iran

Abstract

This paper is concerned with the analysis of correlated responses in insurance data. A Gaussian copula-based regression model is proposed that accounts for associations between the number of automobile and third party claims. Our approach entails specifying underlying latent variables for the responses to indicate the latent mechanisms which generate the count variables. Full likelihood-based inference method is applied for estimation for parametric models to obtain maximum likelihood estimates of the parameters. To illustrate the utility of the models, the proposed methodology is illustrated using some simulation insurance data, with two correlated responses, the count responses of the number of automobile and third party claims. The effect of the risk factor on both responses are also investigated.

Keywords: The Number of Claims, Gaussian Copula, Latent Variable, Insurance Data, Correlated Responses.

[1] f_rezaei@sbu.ac.ir
[2] ehsan_bahrami_samani@yahoo.com
[3] m-ganjali@sbu.ac.ir

1 Introduction

2 One of the essential parts of insurance pricing is modeling insurance claim counts. Typically a comprehensive record of the claim history of their customers is held by insurance companies and they have access to an additional set of personal information. The number of claims reveals the riskiness of the insured. Thus by examining the relation between claim counts and policyholders characteristics, the insurer classifies the policyholders and determines the fair premium according to their risk level. When various types of coverage are tied into one single policy it is common for an insurer to observe claim counts of multiple types from a policyholder. Our goal is to develop bivariate count regression models that accommodate dependency between automobile and third party insurance claims. On the other hand, The separate analysis cannot assess the effect of explanatory variables on both responses. Also, separate analysis gives biased estimates of the parameters, so researchers need to consider a method in which these responses can be modelled jointly.
For analyzing and modeling the number of damages, counting models such as Poisson regression models and negative binomial regressions with cross-sectional responses were used by [9]. [14] introduced a method to estimate correlated discrete random variables with known univariate distribution functions up to some parameters. Various studies have been carried out on the calculation of premiums using multi-dimensional Poisson distributions and negative binomial distributions by [1].

Another method for analyziing correlated count data involves the use of Copulas. This strategy involves the use of diffrenet Copula,for example Gaussian Copula and t Copula ([12], [4] ,and [13]).

[2] calculated the Gaussian Copula From the multivariate standard normal distribution. [5], [10],and [8] discussed Couplas. [6] extended Copula to the case where some of the marginals are a mixture of discrete and continuous components. Copulas is useful where the relevent joint distribution is either not available or difficult to specify but marginal distributions can be specified with confidence.

[7] presented a method for analyzing multivariate count data by Copula modeling. Particularly, he presented techniques for estimating marginal likelihoods and bayes factores in Copula model. [11] considered two different methods of modeling multivariate claim counts using copulas. The first model worked with the discrete count data directly using a mixture of max-id copulas that allows for flexible pair-wise association, tail, and global dependence. The second one employed elliptical copulas to join continuous data while preserving the dependence structure of the original counts.

In this paper, we present new models for bivarite count responses by using Guassian Copula. So far, the researchers did not study models for correlated count responses based on latent variables . Now, here, we present Gussian Copula joint models.

This paper is organized as follows. In Section 2, Gaussian Copula is discussed. In Section 3,We introduce the Gaussian Copula joint model for correlated count responses. Section 4

focuses on insurance claims frequancy data set, simulation studies are used to assure the true model. Finally, in Section 5, the paper discuss with some remarks.

2 Gaussian Copula

[12] introduced a Copula as a function that presents the joint distribution in terms of its marginal. He showed that if H is a bivariate distribution function with margins F and G then there is a joint function such that:

$$H(x, y) = C(F(x), G(y)).$$

In the bivariate case, the approach relates an arbitrary joint distribution $F_{X,Y}$ to its corresponding univariate marginal distribution F_X and F_Y via a copula C as

$$F_{X,Y}(x, y) = C(u, v; \rho),$$

where u and v are respective realization of probability integral transformations $U = F_X(X) \sim uniform[0, 1]$, $V = F_Y(Y) \sim uniform[0, 1]$ and ρ is a dependence parameter measuring dependence between marginals F_X and F_Y.
Gaussian Copulas are an important family which has been used in a variety of applications. The bivariate Gaussian Copula is defined with the standard bivariate normal CDF and has the following form:

$$C_\rho(u_1, u_2; \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\rho)$$

for $u_1, u_2 \in [0, 1]$, where $\Phi(.)$ is the standard normal CDF and $\Phi_2(., .; \rho)$ is the standard bivariate normal CDF, with correlation ρ.
If we assume that the vector variables X have a multivariate Gaussian distribution with correlation matrix R then the copula of X may be represented by

$$C_R^{Ga}(u_1, u_2, \ldots, u_m) = \Phi_R(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_m)),$$

where Φ_R denotes the joint distribution function of a standard d-dimensional normal vector with correlation matrix R, and $\Phi(.)$ is the distribution function of univariate standard normal.

3 Guassian Copula joint Model

Let Y_{ij} indicate count responses (the number of automobile and third party claims) for the ith individual in the jth count for $i = 1, \ldots, nandj = 1, 2$, also let power series distribution be the general form of the probability mass function (pmf) of a random variable (Y_i) with parameter θ_i which can be given as:

$$P(Y_i = k) = \frac{a(k_i)\theta_i^k}{g(\theta_t)}, \qquad y = 0, 1, \ldots, \ \theta_i > 0, \ t = 1, \ldots, T$$

where $a(k_i) > 0$ and $g(\theta_i) = \sum_{n=0}^{\infty} a(k_i)\theta_i^k$ is the normalizing constant. The Poisson, binomial and negative binomial belong to this class. Moreover, we suppose that Y_{ij} takes its values accoring to the latent variable Y_{ij}^* as:

$$Y_{ij} = k_j \, I(\gamma_{j,k_j-1} < Y_{ij}^* < \gamma_{j,k_j}), \; i = 1, ..., n, \;\; j = 1, 2$$

where γ_{j,k_j} determine the discretization of the data, and $\gamma_{j,-1} = -\infty < \gamma_{j,0} < \ldots < \gamma_{j,q-1} < \gamma_{j,q} = \infty$ are cut-points parameters. The latent variable responses (Severity of damages) is modeled as:

$$Y_{ij} \sim PS(\theta_{ij}), \quad q(\theta_{ij}) = X_{ij}^{(1)'}\alpha_j,$$

$$Y_{it}^* = X_{ij}^{(2)'}\beta_j + \varepsilon_{ij},$$

where $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are vectors of the covariates for the ith individual in the jth count and β_j and α_j are vectors of the corresponding unknown regression coefficients. Finally $\epsilon_{ij} \sim N(0, \sigma_1^2)$ fo r$i = 1, \ldots n$.

The Gaussian Copula joint modelling is defined as:

$$\begin{cases} Y_{ij} \sim PS(\theta_{ij}) \;\; \log\theta_{ij} = X_{ij}^{(1)}\alpha_j, \;\; Y_{ij}^* = X_{ij}^{(2)'}\beta_j + \varepsilon_{ij}, \\ \gamma_j, k_j = \Phi^{-1}(P(Y_{ij} = k_j|\theta_{ij})), for j = 1, 2 \end{cases} \tag{3.1}$$

where $(\epsilon_{i1}, \epsilon_{i2}) \sim C_2$ and C_2 is the Gaussian Copula which for $u, v \in [0, 1]$ is defined as:

$$C_2 = \Phi_\rho^2(\Phi^{-1}(u), \Phi^{-1}(v)|\rho),$$

and

$$P(Y_{ij} = k|\theta_{ij}) = \int_{\gamma_{k_t-1}}^{\gamma_{k_t}} \Phi(z|0, 1, \theta_{ij})dz = \Phi(\eta_{k_t-1}) - \Phi(\eta_{k_t}),$$

and,
$v = \frac{Y_{ij}^* - \mu(Y_{ij}^*|b_i^{(1)})}{\mathrm{var}(Y_{ij}^*)}$, $\mu(Y_{ij}^*) = x_{ij}^{'(1)}\beta_j$, and $\mathrm{var}(Y_{ij}^*) = \sigma^2$.
The likelihood function for the model is as:

$$L(\xi) = \prod_{i=1}^{n} P(Y_{i1} = k_1, Y_{i2} = k_2),$$

where $\xi = (\alpha_j, \beta_j, \rho)'$, by using Gaussian Copula we have:

$$P(Y_{i1} = k_1, Y_{i2} = k_2) = P(\gamma_{1,k-1} \leq Y_{i1}^* \leq \gamma_{1,k}, \gamma_{2,k-1} \leq Y_{i2}^* \leq \gamma_{2,k})$$

$=\mathrm{F}_{Y_{i1}^*,Y_{i2}^*}(\gamma_{1,k},\gamma_{2,k}) - F_{Y_{i1}^*,Y_{i2}^*}(\gamma_{1,k},\gamma_{2,k-1}) - F_{Y_{i1}^*,Y_{i2}^*}(\gamma_{1,k-1},\gamma_{2,k}) + F_{Y_{i1}^*,Y_{i2}^*}(\gamma_{1,k-1},\gamma_{2,k-1})$
where

$$F_{Y_{i1}^*,Y_{i2}^*}(y_{i1}^*,y_{i2}^*) = \Phi_2(\Phi^{-1}\{F_{Y_{i1}^*}(y_{i1}^*),\Phi^{-1}\{F_{y_{i2}^*}(y_{i2}^*)|\rho),$$

F_{Yi1} is the marginal distribution of y_{it}, and $\eta_{k_j} = \Phi^{-1}(P(Y_{ij} = k_j|\theta_{ij})$.

The general methods of estimation for the parameters are based on the full likelihood. In this method, the likelihood approach for comutational implementaion needs the log-likelihood which can be maximized by function ″nlminb″ in software R. This function may be used for minimization of a function of parameters. For maximization of a likelihood function one may minimize minus log likelihood function.The function ″nlminb″ uses optimization method of port routine which is given in ″http://netlib.bell-labs.com/cm/cs/cstr/153.pdf″. The function ″nlminb″ uses a sequential quadratic programming (SQP) method to minimize the requested function. The details of this method can be found in [3]. The observed Hessian matrix may be obtained by nlminb function or may be provided by function ″fdHess″.

4 An Insurance Claims Frequency Data Set

To demonstrate the use of the our model, an automobil and third party insurance resposes, a simulation data set is generated. In this data set, claims frequancy and explanatory variables, are obtianed. This simulation study is preformed to evaluate the performance of the model. The simulation study is considered count variables (the number of automobile and third party claims) $Y_j (j = 1, 2)$ to have the power series distribution. The distributions of claims counts Y_j are considered as Poisson distributions. They are generated by their probability functions. The vector of cut points is chosen as $\gamma_j, k = \Phi^{-1}(F(Y_{ij} = k_t|\lambda_{ij}))$ and we consider

$$log(\lambda_{ij}) = \alpha_j^{(0)} X_{ij}^{(2)'} \alpha_j^1$$

where $X_{ij}^{(2)}$ are generated from normal distributions. The initial value for the vector of parameters is chosen as $(\alpha_1^{(0)}, \alpha_2^{(0)}, \alpha_1^{(1)}, \alpha_2^{(1)}) = (0, 0, 1, 1)$.

Also

$$Y_{ij} = \begin{cases} 0 & Y_{ij}^* \leq \gamma_{j,0} \\ 1 & \gamma_{j,0} < Y_{ij}^* \leq \gamma_{j,1} \\ > 2 & \gamma_{j,1} < Y_{ij}^* \end{cases}$$

where (Severity of damages) Y_{ij}^* have a normal distribution with mean $X_{ij}^{2'}\beta_j$ and variance σ^{*2}. Initial values for vector of parameters are chosen as $(\beta_1, \beta_2, \sigma^{*2}) = (1, 1, 1)$. Also, X_{ij}^2s are generated from gamma distributions.

We consider sample size n to be 50, 100, and 1000. In order to ensure the results of the numerical algorithm, the model is fitted by "nlminb" package from software R. Table 1 shows

the result of the simulation study for the model. The parameter estimates are close to the true values of the parameters. Also, the bigger the sample size the smaller the standard error (SE), and as figure 1 shows, the bigger the sample size the smaller mean square error (MSE) which shows the consistency property of our maximum likelihood estimators (MLEs).

Table I. Results of the simulation study for the model.

		n=50		n=100		n=1000	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
$\alpha_1^{(0)}$	0.000	0.259	0.264	0.187	0.185	0.012	0.057
$\alpha_1^{(1)}$	1.000	0.700	0.207	1.117	0.198	0.937	0.049
$\alpha_2^{(0)}$	0.000	0.282	0.212	-0.057	0.191	-0.014	0.063
$\alpha_2^{(1)}$	1.000	0.942	0.274	1.193	0.207	0.969	0.052
β_1	1.000	0.920	0.112	0.928	0.072	1.002	0.022
β_2	1.000	0.926	0.0878	0.926	0.080	1.007	0.024
ρ	0.900	0.865	0.143	0.868	0.136	0.911	0.022

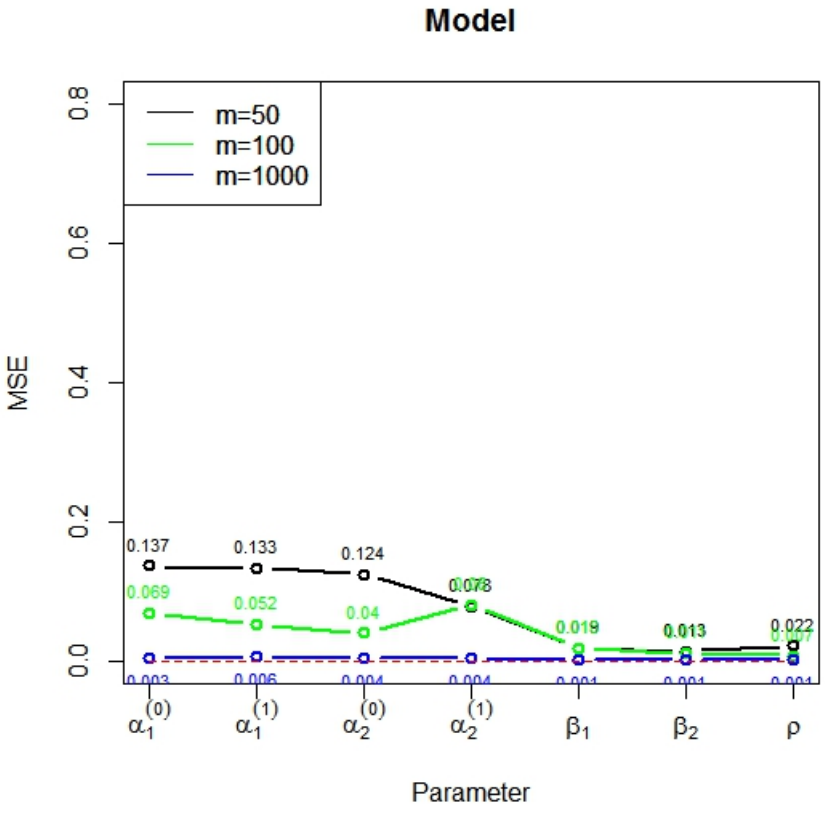

Figure 1: MSE for the model

5 Discussion

Accurate modeling of the correlated claim count responses is one of the essential steps in calculating policy rates. Motivated by the correlated claim count responses of the insurance data set, in this article, we considered alternative approaches to construct correlated count regression models based on Gaussian Copula function. dependency between automobile and third party insurance claims were calculated. In the future studied, we can consider zero-inflated models in the presence of missing data.

References

[1] Jean-Philippe Boucher, Michel Denuit, and Montserrat Guillen, *Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions*, Variance **2** (2008), no. 1, 135–162.

[2] Paul Embrechts, Filip Lindskog, and Alexander McNeil, *Modelling dependence with copulas*, Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich (2001).

[3] Roger Fletcher, *Practical methods of optimization john wiley & sons*, New York **80** (1987).

[4] Maurice J Frank, *On the simultaneous associativity off (x, y) andx+y- f (x, y)*, Aequationes mathematicae **19** (1979), no. 1, 194–226.

[5] Rüdiger Frey, Alexander J McNeil, and Mark Nyfeler, *Copulas and credit models*, Risk **10** (2001), no. 111-114.10.

[6] David Gunawan, Mohamad A Khaled, and Robert Kohn, *Mixed marginal copula modeling*, Journal of Business & Economic Statistics (2018), no. just-accepted, 1–35.

[7] Hee. Esther Lee, *Copula analysis of correlated counts*, Bayesian Model Comparison, Working Paper. University of California, Irvine, CA., 2014, pp. 325–348.

[8] Roger B Nelsen, *An introduction to copulas*, Springer Science & Business Media, 2007.

[9] Danny Samson and Howard Thomas, *Linear models as aids in insurance decision making: the estimation of automobile insurance claims*, Journal of Business Research **15** (1987), no. 3, 247–256.

[10] Arkady Shemyakin and Heekyung Youn, *Copula models of joint last survivor analysis*, Applied Stochastic Models in Business and Industry **22** (2006), no. 2, 211–224.

[11] Peng Shi and Emiliano A Valdez, *Multivariate negative binomial models for insurance claim counts*, Insurance: Mathematics and Economics **55** (2014), 18–29.

[12] M Sklar, *Fonctions de repartition an dimensions et leurs marges*, Publ. inst. statist. univ. Paris **8** (1959), 229–231.

[13] Peter Song, *Multivariate dispersion models generated from gaussian copula*, Scandinavian Journal of Statistics **27** (2000), no. 2, 305–320.

[14] Hans Van Ophem, *A general method to estimate correlated discrete random variables*, Econometric Theory **15** (1999), no. 2, 228–237.

Fifth seminar on
Copula Theory and its Applications
30 & 31 Jan. 2019

Copula Denstiy Estimation by using Legundre Polynomials

Shams, S. [1] Rashidi, H. [2]

Department of Statistics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract

Recently by using contamination families, a new way of modelling dependence has been introduced. In this method a sequence of parametric copulas is considered and in adequate number of steps, accurate approximations for copula densities are obtained. Because it is necessary to balance between model complexity and the number of model parameters, by using the selection model method, the data determine that which aspects are too important to capture into the model. In this paper two main variables in Iranian Households Income and Expenditure survey, are considered and a copula density for those variables is estimated by using selection model.

Keywords: Contamination family, Copula density, Fourier coefficients, Legendre polynomials, Selection model.

[1] s.shams@alzahra.ac.ir
[2]

1 Introduction

The copula approach is a useful method for deriving joint distributions given the marginal distributions. The term copula was introduced by Sklar (1959). Hoeffding established best possible bounds for these functions and studied measures of dependence that are invariant under strictly increasing transformations. Since considering linear correlation in many applications is restricted, other forms of dependence are considered by Embrechts et al. (2003) and Mc Neil et al. (2005). Relationships of copula to the other works have been described in Nelsen (2006). In Biau and Wegkamp (2005) the problem of estimation of copula densities, given a copula density as starting point is discussed. Kallenberg (2008) introduced one approach based on exponential families. Kallenberg (2009) focused on estimating the (unknown) copula density by selection method. In this method, the modelling step consists of an intermediate approach between a parametric family and a non-parametric approach. This is done by considering a sequence of parametric copula models, to yield a sequence of closer and closer approximations to the true copula density. The starting point is a given copula density or a given family of copula densities. There should be a balance between the complexity of the model and the number of parameters involved. To get such a balance, model selection techniques are applied. In this way the data tell us which aspects are the most important ones to capture into our model. The model is kept as simple as possible, but if a more complicated model gives a better fit, it is applied. The penalty in the selection step ensures that only a real improvement is awarded. The unknown parameters within the chosen contamination family are estimated with moment estimators.

This paper is organized as follows. Section 2 deals with some preliminaries. In section 3 the exponential families are reviewed and the decomposition of the total error into the model error and the stochastic error is explained. In Section 4 the contamination families based on Legendre polynomials are reviewed, also this section deals with the model selection problem. The natural way to select the adequate model is to add new parameters as long as a substantial reduction of the model error. A suitable penalty function, depending on the number of observations and the dimension of the model, is a key function. It is shown that within a contamination family with a fixed, but unknown dimension, the selection rule is to choose the 'right' dimension with fast convergence to probability 1. In section 5, for two main variables, Income and Expenditure, in Iranian Households Income and Expenditure Survey, the nearest approximation of copula density using selection method is obtained.

2 Preliminaries

A 2-dimensional **copula** is a function $C : [0,1]^2 \to [0,1]$ with the following properties:
1) For every $u, v \in [0,1]$, $C(0,v) = C(u,0) = 0$;
2) For every $u, v \in [0,1]$, $C(u,1) = u, C(1,v) = v$;

3) For every $(u_1, v_1), (u_2, v_2) \in [0,1] \times [0,1]$ with $u_1 \leq u_2, v_1 \leq v_2$;

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

The theoretical basis of multivariate modelling by copulas is provided by a theorem due to Sklar(1959).

Sklar's Theorem: Let F be a joint distribution function with margins F_1, F_2 which are respectively the cumulative distribution functions of the random variables X_1 and X_2. Then there exists a copula function C such that

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

for every $x_1, x_2 \in \bar{R}$ where $\bar{R}$ represents the extended real line. Conversely if C is a copula and F_1, F_2 are distribution functions then the function F defined a joint distribution function with margins F_1, F_2.

The parametric copula approach ensures a high level of flexibility for modelling, since the specification of the margins F_1 and F_2 can be separated from the specification of the dependence structure through the function C with an underlying parameter θ which governs the intensity of the dependence.

In the case that the bivariate distribution has a density f, and this is available, we have

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)).f_1(x_1).f_2(x_2)$$

where c is the copula density.

The objective of this paper is to estimate the (unknown) copula density c for Iranian Income and Expenditure. In general a natural and very useful way to describe a smooth function on the interval $(0,1)$ is to apply the orthonormal system of Legendre polynomials. This leads for a function z on $(0,1)$ as

$$z(u) = \sum_{r \geq 0} \gamma_r b_r(u)$$

where b_r is the r^{th} Legendre polynomial on $(0,1)$ and γ_r is the r^{th} Fourier coefficient, given by $\gamma_r \leq z, b_r \geq \int_0^1 z(u) b_r(u) du$.

For example the Legendre polynomials $b_0, ..., b_5$ are given by

$$\begin{aligned}
b_0(u) &= 1 \\
b_1(u) &= \sqrt{3}(2u-1) \\
b_2(u) &= \sqrt{5}(6u^2 - 6u + 1) \\
b_3(u) &= \sqrt{7}(20u^3 - 30u^2 + 12u - 1) \\
b_4(u) &= 3(70u^4 - 140u^3 + 90u^2 - 20u + 1) \\
b_5(u) &= \sqrt{11}(252u^5 - 630u^4 + 560u^3 - 210u^2 + 30u - 1)
\end{aligned} \tag{2.1}$$

3 Exponential families

The exponential families are well-known families of parametric models that are used for approximating copula density function. If c_0 is the starting copula density function, the desired copula density is then approximated by

$$c_k(u,v;\theta) = c_0(u,v)\exp\{\sum_{j=1}^{k}\theta_j h_j(u,v) - \psi_k(\theta)\} \tag{3.1}$$

where $h_j(u,v) = b_{r_j}b_{s_j}$, b_{r_j} and b_{s_j} are Legendre polynomials, and $\theta = (\theta_1,\ldots,\theta_k)$ and ψ_k is a normalizing function, given by

$$\psi_k(\theta) = \log\int\int c_0(u,v)\exp\{\sum_{j=1}^{k}\theta_j h_j(u,v)\}dudv \tag{3.2}$$

In order to balance between complexity and the number of parameters, dimension k is determined. Note that c_0 may contain an unknown parameter, which should be estimated as well. In fact, in this way $\log(\frac{c_k}{c_0})$ is approximated by a linear combination of the functions h_j minus a normalizing factor ψ_k to make its integral equal to 1. Exponential families ensure automatically that we get densities such that θ belongs to the natural parameter space

$$\Theta = \left\{\theta : \int\int c_0(u,v)\exp\{\sum_{j=1}^{k}\theta_j h_j(u,v)\}dudv < \infty\right\} \tag{3.3}$$

By considering the Kullback Leibler information and using $argminK(c,c_k(\theta))$ as the appropriate approximation we have,

$$\begin{aligned} K(c,c_k(\theta)) &= E_c\log\left(\frac{c}{c_k(\theta)}\right) = E_c\log c - E_c\log(c_k(\theta)) \\ &= E_c\log c/c_0 - \left\{\sum_{j=1}^{k}\theta_j E_c h_j - \psi_k(\theta)\right\} \\ &= K(c,c_0) - \left\{\sum_{j=1}^{k}\theta_j E_c h_j - \psi_k(\theta)\right\} \\ &= K(c,c_0) - K(c_k(\theta),c_0) + \sum_{j=1}^{k}\theta_j(E_\theta h_j - E_c h_j). \end{aligned} \tag{3.4}$$

It is seen that minimizing $K(c,c_k(\theta))$ is equivalent to maximizing $\sum_{j=1}^{k}\theta_j E_c h_j - \psi_k(\theta)$, which gives the asymptotic version of the maximum likelihood estimator. So, asymptotically the

maximum likelihood estimator chooses that member $c_k(\theta)$ of the exponential family which is closest to the true density c in terms of Kullback Leibler information.
Kallenberg (2008) showed that $c_k(\tilde{\theta})$ is the projection of c into the exponential family with "base" c_0 or

$$K(c, c_0) = K(c, c_k(\tilde{\theta})) + K(c_k(\tilde{\theta}), c_0). \tag{3.5}$$

Where $\tilde{\theta} \in int\Theta$ is a unique point such that $K(c, c_k(\tilde{\theta})) = min\{K(c, c_k(\theta)); \theta \in \Theta\}$
Hence the model error $K(c, c_0)$ is reduced to $K(c, c_k(\tilde{\theta}))$, with a reduction equal to $K(c_k(\tilde{\theta}), c_0)$. Another extra reduction from taking a higher dimension, when going from k to k+1 the model error is occured by an amount $K(c_{k+1}(\tilde{\theta}_{k+1}), c_0) - K(c_k(\tilde{\theta}_k), c_0)$. For the exponential family, the better fit means the smaller model error and the higher dimension or the more parameters have to be estimated. Since parameters estimation in the exponential family is difficult, the idea of contamination family is developed.

4 Contamination families

Just like the exponential family, here the starting point is a copula density c_0, and $c - c_0$ is approximated by a linear combination of the functions $b_r(U)b_s(V)$, hence

$$c_k(u, v) - c_0(u, v) = \sum_{j=1}^{k} \gamma_{r_j s_j} b_{r_j}(u) b_{s_j}(v) \tag{4.1}$$

where γ_{rs} are Fourier coefficients as follows

$$\begin{aligned}\gamma_{rs} &= \int\int \{c(u, v) - c_0(u, v)\} b_r(u) b_s(v) du dv = E_c(b_r(U)b_s(V)) - E_{c_0}(b_r(U)b_s(V)) \\ &= \rho(b_r(U), b_s(V); c) - \rho(b_r(U), b_s(V); c_0)\end{aligned}$$

These coefficients depend on the unknown copula density function c that if it is replaced with empirical copula mass function c_n, then γ_{rs} can be estimated as

$$\hat{\gamma}_{rs} = \frac{1}{n} \sum_{i=1}^{n} b_r(U_i) b_s(V_i) - E_{c_0} b_r(U) b_s(V) \tag{4.2}$$

So when the starting copula density function c_0 belongs to a parametric family, its parameters should be estimated, then we have

$$\hat{c_k}(u, v) = c_0(u, v) + \sum_{j=1}^{k} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v)$$

Kallenberg (2009) showed that the term $\| c - \hat{c_k}(\theta) \|_2^2$ can be written as

$$\begin{aligned} \| c - \hat{c_k} \|_2^2 &= \| c - c_k \|_2^2 + \| c_k - \hat{c_k} \|_2^2 \\ &= \left(\sum_{r,s} \gamma_{rs}^2 - \sum_{j=1}^{k} \gamma_{r_j s_j}^2 \right) + \sum_{j=1}^{k} (\gamma_{r_j s_j} - \hat{\gamma}_{r_j s_j})^2. \end{aligned} \tag{4.3}$$

where $\| f \|_2^2 = \int \int f(u,v) du dv$.
By equation (9) it can be seen that Total Error is decomposed by Model Error and Stochastic Error, or

$$Total\ Error = Model\ Error + Stochastic\ Error$$

The model error $\| c - c_k(\tilde{\theta}) \|_2^2$ expresses how good the contamination family approximates the true density c and the stochastic error $\| c_k(\tilde{\theta}) - c_k(\hat{\theta}) \|_2^2$ is due to estimation.

4.1 Model Selection

In order to obtain parameter estimations in contamination family, the best dimension should be chosen. Suppose m_n be the largest dimension of r and s with n observations, then we have

$$c_0(u,v) + \sum_{r=1}^{m_n} \sum_{s=1}^{m_n} \hat{\gamma}_{rs} b_r(u) b_s(v) \tag{4.4}$$

as the copula density estimation.
For the selection rule, taking all the coefficients $\hat{\gamma}_{rs}$ for $1 \leq r, s \leq m_n$ yields a large estimation error, so we consider only the largest Fourier coefficients and ignore the rest. Therefore, the estimator from (9), is replaced by restricting to the k_n largest among $\hat{\gamma}_{rs}$ with $1 \leq r, s \leq m_n$, yielding

$$\hat{c}(u,v) = c_0(u,v) + \sum_{j=1}^{k_n} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v)$$

with

$$| \hat{\gamma}_{r_1 s_1} | \geq | \hat{\gamma}_{r_2 s_2} | \geq \ldots \geq | \hat{\gamma}_{r_{K_n} s_{K_n}} | .$$

Random variables r_j and s_j depend on the data, and they are not chosen in advance. So how large should we take k_n? The optimal choice depends on c, but c is unknown, hence a data-driven selection of the dimension is taken.
The model error for $\hat{c}_k$ is $\sum_{r,s} \gamma_{rs}^2 - \sum_{j=1}^{k} \gamma_{r_j s_j}^2$. Hence, $\sum_{j=1}^{k} \gamma_{r_j s_j}^2$ should grow sufficiently fast in order to take a higher dimension. For that purpose a penalty is introduced, this penalty is linear in k and decreasing in n. Obviously, we do not know the γ_{rs} and therefore we replace them by $\hat{\gamma}_{rs}$. Classical penalties are for example $n^{-1} \log n$ (Schwarz's rule) or

Table 1: Descriptive Statistics for Income and Expenditure data of Urban households

Descriptive statistics	$Income(10^6 Rials)$	$Expenditure(10^6 Rials)$
$Minimum$	4.914	1.400
Q_1	141.30	124.41
$Median$	207.43	182.18
$Weighted\ \ Mean$	278.78	262.87
Q_3	299.27	270.61
$Maximum$	5787	4424

$2n^{-1}$ (Akaike's criterion). It may be better to take a larger penalty, taking into account the variance of $\hat{\gamma_{rs}}^2$. Kallenberg (2009), introduced a penalty as

$$\Delta_n = n^{-1}(\log n)(\log m_n). \tag{4.5}$$

The estimated copula density now becomes

$$\hat{c}(u,v) = c_0(u,v) + \sum_{j=1}^{\hat{k}} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v). \tag{4.6}$$

5 Copula density estimation for Iranian Households Income and Expenditure

In this section, model selection method is used to estimate copula density for Iranian Households Income and Expenditure (IHIE). The 2015 IHIE survey was carried out by a sample of 18839 households in urban areas and 19340 households in rural areas. The survey target population includes all private and collective settled households in urban and rural areas. A three-staged cluster sampling method with strata is used in the survey. At the first stage, the census areas are classified and selected. At the second stage, the urban and rural blocks are selected and the selection of sample households is done at the third stage. The number of samples is optimized to estimate average annual income and expenditure of the sample household based on the aim of the survey. Income and Expenditure descriptive statistics of urban and rural household are shown in Tables 1 and 2, respectively.

Table 2: Descriptive Statistics for Income and Expenditure data of Rural households

Descriptive statistics	$Income(10^6 Rials)$	$Expenditure(10^6 Rials)$
$Minimum$	3.6	1.23
Q_1	84.52	79.41
$Median$	136.81	124.77
$Weighted\ \ Mean$	161.19	147.26
Q_3	205.58	185.43
$Maximum$	5743	3876

5.1 IHIE copula density estimation with contamination families

By using empirical distributions as marginal distribution estimations for both variables as

$$F_n^X(x) = (n+1)^{-1} \sum_{i=1}^{n} 1(X_i \leq x)$$

the problem is to estimate the unknown copula function. In order to use a few largest Fourier coefficients, the absolute value of the Fourier coefficients are arranged from largest to smallest, and compare with $\sqrt{\Delta_n} = \sqrt{n^{-1} \log n \log m_n}$. By using the sample size of each data set $\sqrt{\Delta_n}$ is calculated, then according to the chosen algorithm of Fourier coefficients, these coefficients are obtained. With several start copula densities as Uniform, Gaussian, Clayton and Frank, as it is shown in Tables 3 and 4, we have several estimations of copula density for rural and urban data sets.

It should be noted that without using this method (selection method) among known copula densities, Frank copula and Clayton copula are the appropriate copulas for Urban and Rural data respectively, here these copulas can be chosen as starting points.

5.2 Investigating performance of the estimated copula function

To check the performance of the estimated copula densities, frequency of data is compared with the estimated probabilities, based on mean absolute relative error (m.a.r.e), $| \frac{\hat{c}}{freq-1} |$, on the same symmetric rectangles ($u = v = 0.25, 0.4$) and asymmetric rectangles ($u = 0.25, v = 0.5; u = 0.5, v = 0.25$) and also the corresponding upper tail rectangles.

As it can be seen from Table 5 and 6 in columns $c_0^U/freq$ the values are far from number one and therefore, it is concluded that when $\hat{k} = 0$ the estimating of dependence with uniform starting copula density function is not good. But by usinf Uniform copula, as starting copula density function, the selection method build better approximations, column $\hat{c_0}^U/freq$ in both Tables 5 and 6 shows these improvements.

Table 3: Results for Urban Data ($\sqrt{\Delta_{18839}} = 0.029$)

c_0	$Uniform$
$\hat{\theta}$	$-$
$\hat{\gamma_{rs}}$	$\hat{\gamma}_{11} = 0.0764, \hat{\gamma}_{22} = 0.0592, \hat{\gamma}_{33} = 0.0429, \hat{\gamma}_{44} = 0.0313$
$\hat{k}$	4
$\hat{c}$	$\hat{c^u}(u,v) = 1 + 0.0764 b_1(u)b_1(v) + 0.0592 b_2(u)b_2(v) + 0.0429 b_3(u)b_3(v) + 0.0313 b_4(u)b_4(v)$
c_0	$Gaussian$
$\hat{\theta}$	0.798
$\hat{\gamma_{rs}}$	$\hat{\gamma}_{33} = 0.0540, \quad \hat{\gamma}_{44} = 0.0738$
$\hat{k}$	2
$\hat{c}$	$\hat{c^G}(u,v) = c_0(u,v;0.798) + 0.0540 b_3(u)b_3(v) + 0.0738 b_4(u)b_4(v)$
c_0	$Frank$
$\hat{\theta}$	1.01
$\hat{\gamma_{rs}}$	$\hat{\gamma}_{22} = 0.1203 \quad \hat{\gamma}_{33} = 0.1804, \quad \hat{\gamma}_{44} = 0.1821$
$\hat{k}$	3
$\hat{c}$	$\hat{c^F}(u,v) = c_0(u,v;6.42) + 0.1203 b_2(u)b_2(v) + 0.1804 b_3(u)b_3(v) + 0.18212 b_4(u)b_4(v)$

Table 4: Results for Rural Data ($\sqrt{\Delta_{19340}} = 0.0286$)

c_0	$Uniform$
$\hat{\theta}$	
$\hat{\gamma_{rs}}$	$\hat{\gamma}_{11} = 0.0738, \quad \hat{\gamma}_{22} = 0.0539, \quad \hat{\gamma}_{33} = 0.0372$
$\hat{k}$	3
$\hat{c}$	$\hat{c^u}(u,v) = 1 + 0.0738 b_1(u)b_1(v) + 0.0539 b_2(u)b_2(v) + 0.0372 b_3(u)b_3(v)$
c_0	$Gaussian$
$\hat{\theta}$	0.754
$\hat{\gamma_{rs}}$	$\hat{\gamma}_{22} = 0.0412, \quad \hat{\gamma}_{42} = 0.0352$
$\hat{k}$	2
$\hat{c}$	$\hat{c^G}(u,v) = c_0(u,v;0.812) + 0.0412 b_2(u)b_2(v) + 0.0352 b_4(u)b_2(v)$
c_0	$Clayton$
$\hat{\theta}$	2.067
$\hat{\gamma_{rs}}$	$\hat{\gamma_{22}} = 0.0699$
$\hat{k}$	1
$\hat{c}$	$\hat{c^C}(u,v) = c_0(u,v;2.067) + 0.0699 b_2(u)b_2(v)$

Table 5: The frequencies and approximations on different rectangles for Urban data

RECTANGLES	$freq$	$c_0^U/freq$	$c_0^G/freq$	$c_0^F/freq$	$\hat{c}^U/freq$	$\hat{c}^G/freq$	$\hat{c}^F/freq$
$(0, 0.25) \times (0, 0.25)$	0.1726	0.362	0.977	0.983	0.986	0.980	1.022
$(0, 0.4) \times (0, 0.4)$	0.2985	0.536	1.006	1.028	0.985	0.997	1.009
$(0, 0.25) \times (0, 0.5)$	0.2290	0.546	1.011	1.003	0.991	0.991	1.001
$(0, 0.5) \times (0, 0.25)$	0.2296	0.544	1.009	1.029	1.090	1.001	1.004
$(0.75, 1) \times (0.75, 1)$	0.1652	0.378	1.021	1.016	1.044	1.015	1.019
$(0.6, 1) \times (0.6, 1)$	0.2950	0.542	1.018	1.037	1.009	1.004	1.021
$(0.75, 1) \times (0.5, 1)$	0.2286	0.546	1.014	1.044	0.992	1.002	1.012
$(0.5, 1) \times (0.75, 1)$	0.2247	0.556	1.031	1.022	1.003	1.013	1.008
$m.a.r.e.$		0.489	0.017	0.024	0.024	0.008	0.012

Table 6: The frequencies and approximations on different rectangles for Rural data

RECTANGLES	$freq$	$c_0^U/freq$	$c_0^G/freq$	${c_0}^C/freq$	$\hat{c}^U/freq$	$\hat{c}^G/freq$	$\hat{c}^C/freq$
$(0, 0.25) \times (0, 0.25)$	0.1739	0.359	0.921	1.128	0.917	0.981	1.058
$(0, 0.4) \times (0, 0.4)$	0.2959	0.541	0.979	1.059	0.988	0.994	1.042
$(0, 0.25) \times (0, 0.5)$	0.2271	0.550	0.993	1.031	0.995	0.994	1.031
$(0, 0.5) \times (0, 0.25)$	0.2303	0.543	0.979	1.059	0.987	0.977	1.052
$(0.75, 1) \times (0.75, 1)$	0.1516	0.412	1.056	0.857	0.979	1.008	0.982
$(0.6, 1) \times (0.6, 1)$	0.2859	0.560	1.013	0.970	1.008	1.019	0.985
$(0.75, 1) \times (0.5, 1)$	0.2223	0.562	1.014	0.988	1.021	1.013	0.992
$(0.5, 1) \times (0.75, 1)$	0.2199	0.568	1.025	0.950	1.014	1.024	0.968
$m.a.r.e.$		0.490	0.029	0.0640	0.022	0.015	0.032

Table 7: The 0.99 and 0.95 estimated and real quantiles for Urban data

probability	*quantile*	*expectation number*	*real number*
0.99	0.9654	189	182
0.95	0.9421	942	940

Table 8: The 0.99 and 0.95 estimated and real quantiles for Rural data

probability	*quantile*	*expectation number*	*real number*
0.99	0.9791	194	190
0.95	0.8753	967	970

Finally, in both Tables 5 and 6, columns $\hat{c}^G/freq$, based on m.a.r.e., it can be seen that Gaussian copula as starting copula, yields appropriate estimations for copula density for Urban and Rural data.

For more investigating of this new method, the quantile 0.99 and 0.95 is obtained by using $\hat{c}^G$, and the values have been compared with the true values (from the IHIE). The results are shown in Tables 7 and 8. In both tables, the expectation numbers and the real numbers are close to each other.

References

[1] Biau, G., Wegkamp, M., 2005. A Note on minimum distance estimation of copula densities. Statistics & probability Letters 73, 105-114.

[2] Emberchts, P., Lindskog, F., McNeil, A., 2003. Modelling dependence with copulas and applications to risk managements. Rachev, S.T. (Ed.), Handbook of Heavy Tailed Distributions in Finance. Elsevier, Amsterdam, 329-384.

[3] Kallenberg, W.C.M., 2008. Modelling Dependence. Insurance: Mathematics and Economics, 2008, vol. 42, issue 1, 127-146

[4] Kallenberg, W.C.M. 2009, Estimating copula densities, using model selection techniques. Journal of Insurance: Mathematics and Economics 45 209-223.

[5] McNeil, A., Frey, R., Emberchts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton.

[6] Nelsen, R.B., 1999. An Introduction to Copulas. Lecture Notes in Statistics, 139. Springer-Verlag, New York.

[7] Sklar, A., 1959. Fonctions de repartition a n dimensions et leurs marges. Punl. Inst. Statist. Univ. Paris 8 229-231.

[8] Sklar, A., 1996. Random variables, distribution functions, and copulas- a personal look backward and forward. In Distributions with Fixed Marginals and Related Topics (L. Ruschendorf, B. Schweizer and M.D. Taylor, eds). 1-14, Lecture notes monograph series 28, Institute of Mathematical Statistics, Hayward, CA.