Article

Statistical Modeling of Insurance Data via Vine Copula

Indranil Ghosh and Dalton Watts

[1] University of North Carolina, Wilmington; ghoshi@uncw.edu
[2] University of North Carolina, Wilmington; dow3346@uncw.edu
* Correspondence: ghoshi@uncw.edu ; Tel.: +1 -910-962-7644

Abstract: Copulas are useful tools for modeling the dependence structure between two or more variables. Copulas are becoming a quite flexible tool in modeling dependence among the components of a multivariate vector, in particular to predict losses in insurance and finance. In this article, we study the dependence structure of some well-known real life insurance data (with two components mainly) and subsequently identify the best bivariate copula to model such a scenario via VineCopula package in *R*. Associated structural properties of these bivariate copulas are also discussed.

Keywords: Bivariate Copula; Measures of association; Dependence modeling; Kendall's τ; Blomqvist's β

1. Introduction

Over the last decade or so, there has been a growing interest in constructing various bivariate and multivariate distributions and study its dependence structure. For an excellent survey on this, an interested reader is suggested to see Balakrishnan et al. (2009) and the references therein. Of late, copula based methods of construction have also gained a considerable amount of attention, mainly due to its analytical tractability in the sense of discussing dependence structure between two dependent random variables. These days, there is a growing trend to analytically evaluate the dependence structure between two or more variables. In general, there are two approaches: (i) setting up a functional relation between the variables, (ii) specifying the joint distribution of the variables. The second approach is much more general as compared to the first one, and calls for a different types of mechanism which eliminates the effect of univariate marginals which has nothing to do with the dependence structure, as it has found in numerous occasions that the effect of marginals indeed distort the original dependence structure between two or more concomitant random variables. In an effort to identify a class flexible bivariate and/or multivariate distributions, copula modeling has recently become increasingly well-known in numerous fields of application. For a detailed study on copula and associated bivariate (as well as multivariate) dependence based on copula theory include the books by Joe (1997) and Nelsen (2006). The seminal theorem, which constitutes the important role of copulas for describing dependence in statistics, is the theorem of Sklar (1959). It establishes the link between multivariate distribution functions and their univariate margins. We state this theorem at first. Let F be the p-dimensional distribution function of the random vector $\underline{X} = \left(X_1, \cdots, X_p\right)^T$ with margins $F_1, \cdots, F_p$. Then there exists a copula C such that for all $\underline{x} = \left(x_1, \cdots, x_p\right)^T \in [-\infty, \infty]^p$,

$$F\left(\underline{x}\right) = C\left(F_1(x_1), \cdots, F_p(x_p)\right). \tag{1}$$

Note that C is unique if $F_1, \cdots, F_p$ are continuous. Conversely, if C is a copula and $F_1, \cdots, F_p$ are distribution functions, then the function F defined by (1) is a joint distribution function with margins

$F_1, \cdots, F_p$. Precisely, C can be interpreted as the distribution function of a p- dimensional random variable on $[0,1]^p$ with uniform margins. Associated densities will be denoted by a lower case c. In addition, the random variables $X_1, \cdots, X_p$ will be assumed to be continuous in the following.
The utility of Sklar's theorem is that the modeling of the marginal distributions can be conveniently and efficiently separated from the dependence modeling in terms of the copula. Interestingly, the major task that lie in practical applications is how to identify this copula. For the bivariate case, a rich collection of copula families is available and well-investigated (see, for details, Joe 1997; Nelsen 2006). However, in arbitrary dimension (precisely, $p \geq 3$), the choice of adequate families is rather limited. Well-known multivariate copulas such as the multivariate Gaussian or Student-t as well as exchangeable Archimedean copulas lack the flexibility of accurately modeling the dependence among larger numbers of variables. Generalizations of these offer some improvement, but typically become rather obscure in their structure and consequently induce other limitations such as parameter restrictions. On the other hand, Vine copulas do not suffer from any of these problems. Initially proposed by Joe (1996) and developed in more detail in Bedford and Cooke (2001, 2002) and in Kurowicka and Cooke (2006), Vines are a flexible graphical model for describing multivariate copulas built up using a cascade of bivariate copulas, so-called pair-copulas. Such pair-copula constructions decompose a multivariate probability density into bivariate copulas, where each pair-copula can be chosen independently from the others. This allows for a enormous flexibility in dependence modeling. In particular, asymmetries and tail dependence can be taken into account as well as (conditional) independence to build more parsimonious models. Therefore, Vines combine the advantages of multivariate copula modeling, that is separation of marginal and dependence modeling, and the flexibility of bivariate copulas. Their statistical breakthrough was due to Aas et al.(2009) who described statistical inference techniques for the two classes of canonical (C-) and D-Vines. C- and D-Vine copulas have been very successful in many applications, mainly, but not exclusively, in risk management in finance and insurance, see, e.g., Schirmacher and Schirmacher (2008), Chollete et al. (2009), Heinen and Valdesogo (2009), de Melo Mendes et al.(2010), Czado et al.(2012), and Nikoloulopoulos et al.(2012). Bayesian approaches are followed by Min and Czado (2010), Min and Czado (2011), Smith et al.(2010), and Hofmann and Czado (2010). Most recent works on the Vine methodology can be found in Czado (2010) and Kurowicka and Joe (2011), which includes further applications and theory.

In this article, we consider the application of Vine copulas (in two dimension) for several types of insurance data which are asymmetric in nature on utilizing the Vine Copula package in *R*. It appears that the resultant most appropriate bivariate copulas are members of the *C* and *D*-Vine copulas and among them couple of them are Archimedean as well. This is a follow up article of Ghosh and Ray (2016) in which the authors discussed some bivariate Kumaraswamy types of copulas with applications in risk management. The paper is organized as follows. In Section 2, we discuss some basic definitions and useful preliminaries on copula and Vine copula theory. In section 3, we discuss in details three different data sets and subsequently fitting an appropriate bivariate copula to each of them. In section 4, we discuss some useful structural properties of these copulas. Some concluding remarks are made in section 5. Finally, the associated *R* codes (for illustrative purposes only) are given in appendix at the end.

2. Bivariate and Vine Copulas: Preliminaries

We begin this section with some basic properties of a copula. For details on this, see Nelsen (1999, 2006).

Definition 1. A copula is a function C whose domain is the entire unit square with the following properties:

1. $C(u,0) = C(0,v) = 0$, for all $(u,v) \in [0,1]$.
2. $C(u,1) = C(1,u) = u$, for all $(u,v) \in [0,1]$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

3. $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$, for all $(u_1, v_1, u_2, v_2) \in [0,1]$. for every $u_1 \leq u_2, \quad v_1 \leq v_2$.

Sklar (1973) established that for any bivariate distribution function, say,$F_{XY}()$, can be represented as a function of its marginals, say, $F_X()$ and $F_Y()$, by using a two dimensional copula $C(.,.)$ in the following way:

$$F_{XY}(x, y) = C\left(F_X(x), F_Y(y)\right).$$

If $F_X()$ and $F_Y()$ are absolutely continuous, then the associated copula C is unique. Also, $C(u, v)$ is ordinally invariant, which implies that $\delta(x)$ and $\Phi(y)$ are strictly increasing functions, the copula of $(\delta(X), \Phi(Y))$ is also that of (X, Y). Therefore, if each marginal of $F_{XY}(x, y)$ is absolutely continuous, then by selection of $\delta(x) = F_X(x)$ and $\Phi(y) = F_Y(y)$, we can say that every copula is a distribution function whose marginals are uniform on the interval $[0, 1]$. Consequently, it represents the dependence structure between two variables by eliminating the influence of the marginals and hence of any monotone transformation on the marginals.

Next, we briefly discuss the concept of Vine copula here. For a detailed study, one is referred to Joe (1997) and Brechman et al. (2013). Vines are a graphical representation to stipulate so-called pair copula constructions (PCC, henceforth, in short) as introduced by Aas et al. (2009). We begin our discussion in this context by motivating the PCC in four dimensions. For this, let $Y = (Y_1, Y_2, Y_3, Y_4)^T \sim F$ with marginal cumulative distribution functions (c.d.f) F_i, $i = 1, 2, 3, 4$ and the corresponding densities (p.d.f.) f_i, $i = 1, 2, 3, 4$ respectively. By recursive conditioning we can write

$$f\left(y_1, y_2, y_3, y_4\right) = f_1(y_1) f_2\left(y_2|y_1\right) f_3\left(y_3|y_1, y_2\right) f_4\left(y_4|y_1, y_2, y_3\right). \tag{2}$$

Then, from Aas et al. (2009), the associated C-vine and D-vine copula density, respectively, will be

$$\begin{aligned} f\left(y_1, y_2, y_3, y_4\right) &= \prod_{i=1}^{4} f_i(y_i) c_{12}\left(F_1(y_1), F_2(y_2)\right) c_{13}\left(F_1(y_1), F_3(y_3)\right) c_{14}\left(F_1(y_1), F_4(y_4)\right) \\ &\times c_{23|1}\left(F_{2|1}(y_2|y_1), F_{3|1}(y_3|y_1)\right) c_{24|1}\left(F_{2|1}(y_2|y_1), F_{4|1}(y_4|y_1)\right) \\ &\times c_{34|12}\left(F_{3|12}(y_3|y_1, y_2), F_{4|12}(y_4|y_1, y_2)\right). \end{aligned} \tag{3}$$

$$\begin{aligned} f\left(y_1, y_2, y_3, y_4\right) &= \prod_{i=1}^{4} f_i(y_i) c_{12}\left(F_1(y_1), F_2(y_2)\right) c_{23}\left(F_2(y_2), F_3(y_3)\right) c_{34}\left(F_3(y_3), F_4(y_4)\right) \\ &\times c_{13|2}\left(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2)\right) c_{24|3}\left(F_{2|3}(y_2|y_3), F_{4|3}(y_4|y_3)\right) \\ &\times c_{14|23}\left(F_{1|23}(y_1|y_2, y_3), F_{4|23}(y_4|y_2, y_3)\right). \end{aligned} \tag{4}$$

Note that, since the decomposition in (3) and in (4) are not unique, there exist many such iterative PCCs. To classify them into a separate class, Bedford and Cooke (2001, 2002) introduced the graphical model called vine. It is also discussed in detail in Kurowicka and Cooke (2006) and Kurowicka and Joe (2011). Vines arrange the $p(p-1)/2$ pair-copulas of a p-dimensional PCC in $(p-1)$ linked trees (acyclic connected graphs with nodes and edges). In the first C-vine tree, the dependence with respect to one particular variable, the first root node, is modeled using bivariate copulas for each pair. Conditioned on this variable, pairwise dependencies with respect to a second variable are modeled, the second root node. In general, a root node is chosen in each tree and all pairwise dependencies with respect to this node are modeled conditioned on all previous root nodes, i.e., C-vine trees have a star structure. This gives the following decomposition of a multivariate density, the C-vine density without loss of generality with root nodes $1, \cdots, p$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

$$f(\underline{y}) = \prod_{j=1}^{p} f_j(y_j) \times \left[\prod_{i=1}^{p-1} \prod_{j=1}^{p-i} c_{i,i+j|1:(i-1)} \left(F\left(y_i | y_1, \cdots, y_{i-1}\right), F\left(y_{i+j} | y_1, \cdots, y_{i-1}\right) \right) \right]. \quad (5)$$

where $f_j(), \quad j = 1, \cdots, p$ denote the marginal densities and $c_{i,i+j|1:(i-1)}$ bivariate copula densities. In this case, the outer product runs over the $p-1$ trees and root nodes i, while the inner product refers to the $p-i$ pair-copulas in each tree $i = 1, \cdots, p-1$. Our four-dimensional example can be interpreted as a C-vine with X_1 as first root node. A more in depth discussion of the C-vine construction and its likelihood can be found in Aas et al. (2009) and in Czado et al. (2012). Similarly, D-vines are also constructed by choosing a specific order of the variables. In this case, in the first tree, the dependence of the first and second variable, of the second and third, of the third and fourth, and so on, is modeled using pair-copulas. Consequently, the associated D-vine density which also conveniently decomposes a p-dimensional density without loss of generality with root nodes $1, \cdots, p$

$$f(\underline{y}) = \prod_{j=1}^{p} f_j(y_j) \times \left[\prod_{i=1}^{p-1} \prod_{j=1}^{p-i} c_{i,i+j|(j+1):(j+i-1)} \left(F\left(y_j | y_{j+1}, \cdots, y_{j+i-1}\right), F\left(y_{i+j} | y_{j+1}, \cdots, y_{j+i-1}\right) \right) \right]. \quad (6)$$

Therefore, by allowing arbitrary bivariate copulas for each pair-copula term in the decompositions (5) and (6), the multivariate copulas obtained from C and D-vine structures, so-called C and D- vine copulas, constitute very flexible models, since bivariate copulas can easily accommodate complex dependence structures such as asymmetric dependence or strong joint tail behavior (for details, see Joe et al. (2010)). For the p-dimensional D-vine, the pairs at level 1 are $i, i+1$, for $i = 1, \cdots, p-1$, and for level ℓ $(2 < \ell < p)$, the (conditional) pairs are $i, i+\ell | i+1, \cdots, i+\ell-1$ for $i = 1, \cdots, p-\ell$. For the p-dimensional C-vine, the pairs at level 1 are $1, i$, for $i = 2, \cdots, p$, and for level ℓ $(2 < \ell < p)$, the (conditional) pairs are $\ell, i | 1, \cdots, \ell-1$ for $i = \ell+1, \cdots, p$. Consequently, for the D-vine, conditional copulas are specified for variables i and $i+\ell$ given the variables indexed in between; and for the C-vine, conditional copulas are specified for variables ℓ and i given those indexed as 1 to $\ell-1$.

In this article, from the fitted most appropriate bivariate copulas (Section 3), it appears that they are special members of the C-vine and D-vine copulas. The structural properties of these specific members of the C and D-vine copulas are discussed in section 4.

2.1. Dependence structures

It is noteworthy to mention that copulas are instrumental for understanding the dependence between random variables. With them we can separate the underlying dependence from the marginal distributions. It is well known that a copula which characterizes dependence is invariant under strictly monotone transformations, subsequently a better global measure of dependence would also be invariant under such transformations. Among other dependence measures, Kendall's τ and Spearman's ρ are invariant under strictly increasing transformations, and, as we will see in the next, they can be expressed in terms of the associated copula.

- **Kendall's τ:** Kendall's τ measures the amount of concordance present in a bivariate distribution. Suppose that (X, Y) and $(\tilde{X}, \tilde{Y})$ are two pairs of random variables from a joint distribution function. We say that these pairs are concordant if large values of one tend to be associated with large values of the other, and small values of one tend to be associated with small values of the other. The pairs are called discordant if large goes with small or vice versa. Algebraically we have concordant pairs if $(X - \tilde{X})(Y - \tilde{Y}) > 0$ and discordant pairs if we reverse the inequality. The formal definition is:

$$\tau(X, Y) = P\left\{ \left(X - \tilde{X}\right)\left(Y - \tilde{Y}\right) > 0 \right) - \left(X - \tilde{X}\right)\left(Y - \tilde{Y}\right) < 0 \right) \},$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

where $(\tilde{X}, \tilde{Y})$ is an independent copy of (X, Y).

Let X and Y be continuous random variables with copula C. Then Kendall's τ is given by

$$\tau(X,Y) = 4 \iint_{[0,1]^2} C(u,v) dC(u,v) - 1. \tag{7}$$

- **Spearman's ρ:** Let X and Y be continuous random variables with copula C. Then Spearman's ρ_s is given by

$$\rho_s = 12 \iint_{[0,1]^2} uv dC(u,v) - 3. \tag{8}$$

Alternatively, ρ_s can be written as $\rho_s = 12 \int_0^1 \int_0^1 [C(u,v) - uv]\, dudv$. Also, as mentioned earlier, one can equivalently show that $\rho_s(U,V) = \rho\left(F_1(X), F_2(Y)\right)$.

- **Tail dependence property:** Let X and Y are two continuous r.v's with $X \sim F$, and $Y \sim G$. The upper tail dependence coefficient (parameter) λ_U is the limit (if it exists) of the conditional probability that Y is greater than 100α th percentile of G given that X is greater than the 100α th percentile of F as α approaches 1.

$$\lambda_U = \lim_{\alpha \uparrow 1} P\left(Y > G^{-1}(\alpha) | X > F^{-1}(\alpha)\right). \tag{9}$$

If $\lambda_U > 0$, then X and Y are upper tail dependent and asymptotically independent otherwise. Similarly, the lower tail dependence coefficient is defined as

$$\lambda_L = \lim_{\alpha \downarrow 0} P\left(Y \leq G^{-1}(\alpha) | X \leq F^{-1}(\alpha)\right). \tag{10}$$

Let, C be the copula of X and Y. Then, equivalently we can write

$\lambda_L = \lim_{u \downarrow 0} \frac{C(u,u)}{u}$ and $\lambda_U = \lim_{u \downarrow 0} \frac{\tilde{C}(u,u)}{u}$,

where $\tilde{C}(u,u)$ is the corresponding joint survival function given by

$$\tilde{C}(u,u) = 1 - 2u + C(u,u).$$

- **Blomqvist's beta:** Suppose that $\tilde{X}_n$ and $\tilde{Y}_n$ be the medians of the samples $X_1, , \cdots, X_n$ and $Y_1, , \cdots, Y_n$ respectively. In order to summarize information about the dependence between X and Y, Blomqvist (1950) suggested dividing the $x - y$ plane into four regions by drawing the lines $x = \tilde{\tilde{X}}_n$ and $y = \tilde{\tilde{Y}}_n$ and comparing the following quantities:

  - n_1 : the number of points lying in either the lower left quadrant or the upper right quadrant;
  - n_2 : the number of points in either the upper left quadrant or the lower right quadrant.

Consequently, the definition of β_n, which is equivalently called Blomqvist's beta, is given by

$$\beta_n = \frac{n_1 - n_2}{n_1 + n_2} = -1 + 2\frac{n_1}{n_1 + n_2}.$$

If n is even, then no sample point falls on either of the lines $x = \tilde{\tilde{X}}_n$ and $y = \tilde{\tilde{Y}}_n$ and it follows that both n_1 and n_2 are even. If n is odd, however, then either one or two sample points lie on the lines defined by the sample medians. In the case of a single point lying on a median, Blomqvist (1950) proposed not to count the point altogether. In the latter case, one point has to fall on each line: one of them is assigned to the quadrant touched by the two points, and the other is not counted. This allows both n_1 and n_2 to remain even. The population analogue of β_n, is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

$$\beta = P\left[(X-\tilde{x})(Y-\tilde{y})>0\right] - Pr\left[(X-\tilde{x})(Y-\tilde{y})<0\right],$$

where $\tilde{x}$ and $\tilde{y}$ denote the population medians of X and Y, respectively. Next, on using the facts that

–

$$\begin{aligned} P\left[(X-\tilde{x})(Y-\tilde{y})>0\right] &= P\left[(X-\tilde{x})>0,(Y-\tilde{y})>0\right] \\ &\quad + P\left[(X-\tilde{x})<0,(Y-\tilde{y})<0\right]; \end{aligned}$$

and $P\left[X>\tilde{x},Y>\tilde{y}\right] = Pr\left[X<\tilde{x},Y<\tilde{y}\right]$;

– From the fundamental Sklar's (1959) theorem $H(x,y) = C\left(F(x),G(y)\right)$;

one can write

$$\beta = 4C\left(\frac{1}{2},\frac{1}{2}\right) - 1. \tag{11}$$

As β is only a function of C, it is possible to write it in terms of $\underline{\alpha}$ whenever $C \in C_{\underline{\alpha}}$, where $\underline{\alpha}$ is the set of parameters associated with the copula C.

- **Left-Tail decreasing property and Right- Tail increasing property:** Nelson (1999) showed that $X(Y)$ is left tail decreasing i.e., $LTD(Y|X)$ and $LTD(X|Y)$ if and only if for all u, u', v, v' such that $0 < u \leq u' \leq 1$ and $0 < v \leq v' \leq 1$, if $\frac{C(u,v)}{uv} \geq \frac{C(u',v')}{u'v'}$. For an alternative criteria, see Nelsen (2006)(page 197,Theorem 5.2.12 and Corollary 5.2.11).

Note that the associated dependence measures for all the fitted bivariate copulas are provided in Table 1.

3. Application to Insurance Data

3.1. Data and Variable Selection

All of the data sets referred to in this paper are found in the Computational Actuarial Science collection and are accessible through the "CASdatasets" package in R. Additonally, we used the "VineCopula" package for our code in order to find the best fit model for each pair of variables in each data set used. The "VineCopula" package takes the selected variables and finds the best copula model from the families available in the package. This choice of copula is based on test diagnostics such as AIC, BIC, and the log-likelihood value. The next section details how the variables from each data set were selected and the results of the different models.

3.1.1. Data Set 1 (Height and Weight Analysis)

This dataset contains the gender (M/F), weight, height, reported weight, and reported height of 200 individuals. Gender was not considered because it is a qualitative categorical variable. Upon further inspection into the data set, it was discovered that reported weight and height possessed missing data values. Consequently, we did not consider those variables in our model. The following variables of interest were selected:

1. X_1: Height-The height of the participant (cm).
2. X_2: Weight-The weight of the participant (Kg).

3.1.2. Data Set 2 (Australian Automobile Claim Data)

This dataset records the number of third party claims in a 12 month period between 1984 and 1986 in each of 176 local government areas in New South Wales, Australia. Additionally, the data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

set includes the name of the local government, the number of third party claims filed, the number of people killed or injured in automobile accidents, the population size, and the population density. Australia is historically known for its low population density. This is due to extreme climate of the continent. With this in mind, we decided to include the population size of each city in New South Wales as opposed to the population density because the density is skewed by the lack of inhabitants in Australia. For this dataset we plan to study dependence measure among 4 different variables in pairwise comparison structure. We argue that the selection of these pairwise comparisons are legitimate in nature. The table below provides a key for the abbreviations we will use for each variable.

Table 1. Variable Name Key

Abbreviation	Variable
ACC	Number of Accidents
TPC	Number of Third Party Claims filed
K/I	Number of people killed or injured in an accident
Pop	Population of the area

Model 1 (AUS 1)

The first pair of variables that were selected were the number of accidents and the population size. These were chosen because we expect more accidents to occur in regions with a higher population relatively speaking.

1. X_1: ACC
2. X_2: Pop

Model 2 (AUS 2)

A third party claim is a claim filed by someone other than the insured and their insurance company. If a driver's negligence results in the injury or death of another driver, the affected party or their family have the ability to file a claim against the guilty driver's insurance company. We decided to measure the dependence between this variable and the population of a given region in Australia because one would expect a larger volume of third party claims to be filed in regions with higher populations.

1. X_1: TPC
2. X_2: Pop

Model 3 (AUS 3)

Next, the dependence of a region's population and the number of people killed or injured in an accident were considered. Once we discovered that there was a strong dependence relationship between the number of third party claims and the population size of a region, we realized that since third party claims are a result of accidents with injuries involved, the number of people killed or injured could be greater in higher populated areas where more third party claims are filed.

1. X_1: Pop
2. X_2: K/I

Model 4 (AUS 4)

In this model, we decided to measure the level of dependence between the number of people injured or killed in an automobile accident and the corresponding number of third party claims filed. As defined above in model 2, third party claims are filed in the event of an accident in which other drivers suffer injury from the negligence of another. While injury and death are not exclusive to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

third party, we found a positive trend in the scatter plot of these two variables. Hence, we chose to fit a copula to these two concomitant variables.

1. X_1: K/I
2. X_2: TPC

3.1.3. Data Set 3 (Swedish Motor Insurance Data)

This data set represents the insurance information of 2,182 motorists collected by the Swedish Committee on the Analysis of Risk Premium in 1977. It consists of the number of kilometres driven by a motorist (grouped into 5 categories), the geographical zone of a vehicle (grouped into 7 categories), the bonus variable (grouped into 7 categories), the make of the vehicle, the number of years that a motorist has been insured, the number of claims a motorist has filed, and the sum of the payments made by a motorist. We excluded the geographic zone and make of the vehicle variables from our consideration because while they are quantitatively defined, they describe qualitative variables and do not have a defined ordering. Due to the way the kilometres variable has been defined, we were unable to come up with a model that showed a large amount of dependence, so the results of that model have been excluded from this paper. Instead, we chose to study the dependence and subsequently search for a best possible bivariate copula model with the following variables of interest:

1. X_1: Insured (Number of years a motorist has been insured)
2. X_2: Claims (Sum of claim payments)

The tables below detail the results of each model. Note that all of these computations were performed in R.

Table 2. Level of dependence between model variables

Data set	X_1	X_2	Kendall's Tau	Spearman's Rho
Davis	Height	Weight	0.6157	0.7954
AUS 1	ACC	Population	0.8123	0.9452
AUS 2	TPC	Population	0.8078	0.9479
AUS 3	K/I	Population	0.7981	0.9373
AUS 4	K/I	TPC	0.8372	0.9611
Swedish Motor	Policy Holder Years	Sum of Payments	0.7411	0.9030

Table 3. Model Diagnostics and Goodness of Fit Statistics

Data set	Best Fitted Copula	Parameter Estimates	AIC	BIC	Log Likelihood
Davis	Gaussian	(0.80)	-192.01	-188.71	97
AUS 1	Frank	(18.42)	-377.38	-374.21	189.69
AUS 2	Frank	(18.33)	-376.58	-373.41	189.29
AUS 3	Tawn 1	(5.01,0.95)	-373.11	-366.76	188.55
AUS 4	Student t	(0.96,4.61)	-442.33	-435.99	223.16
Swedish Motor	BB6	(1.59,2.81)	-4095.96	-4084.58	2049.98

Table 2 outlines the level of concordance between each pair of variables in each model. When two variables are concordant, this means that higher values of one variable are associated with higher values of the other and vice versa for lower values. If these coefficients are closer to 0, this indicates low dependence or even independence. Conversely, if these coefficients are closer to 1, it tells us that the variables are dependent upon one another. From Table 2, we see that each pair of variables exhibit a strong level of dependence since the concordance coefficients are close to 1. Table 3 represents various model diagnostics along with parameter estimates corresponding to the best fitted bivariate copula.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

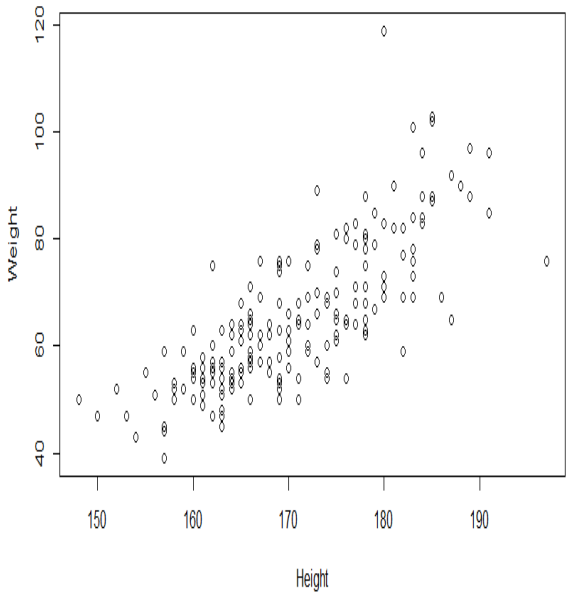


Figure 1. Height vs. Weight

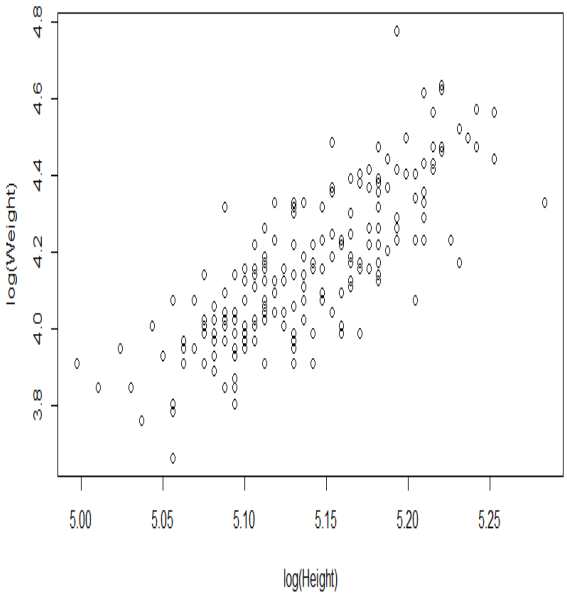


Figure 2. Height vs. Weight (log scale)

We expect that the AIC and BIC to be minimal and the log-likelihood to be maximal. Each copula shown in Table 3 represents the best fit for the pair of variables that were being tested according to the AIC, BIC, and log-likelihood criteria.

3.2. Scatterplots

In this section, we provide plots of the raw data and compare it to the variables on the logarithmic scale. The logarithmic plots are provided in response to the skewness of the original data values. In addition, CDF and PDF plots are also provided corresponding to best fitted bivariate copulas mentioned in Table 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

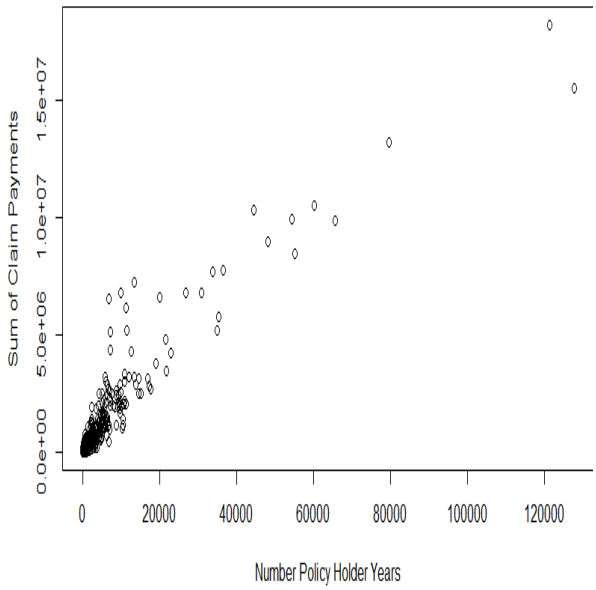


Figure 3. Policy Years vs. Payments

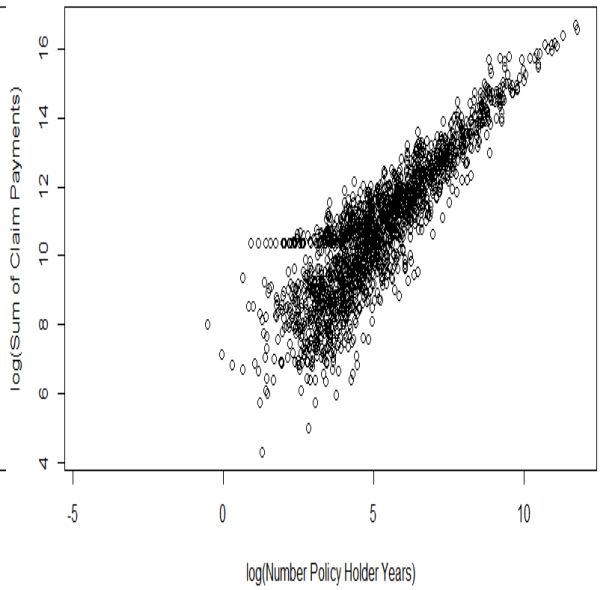


Figure 4. Policy Years vs. Payments (log scale)

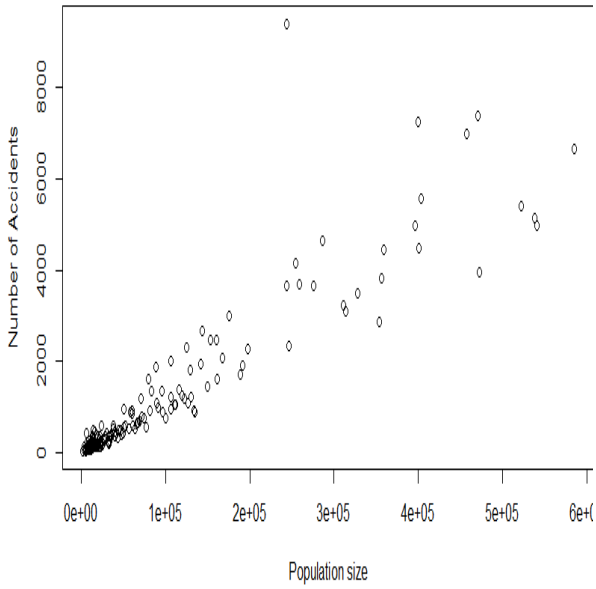


Figure 5. Pop vs. Acc

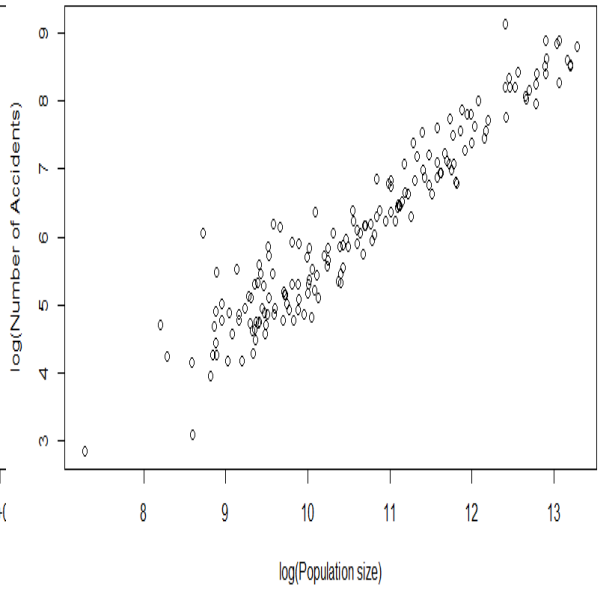


Figure 6. Pop vs. Acc (log scale)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

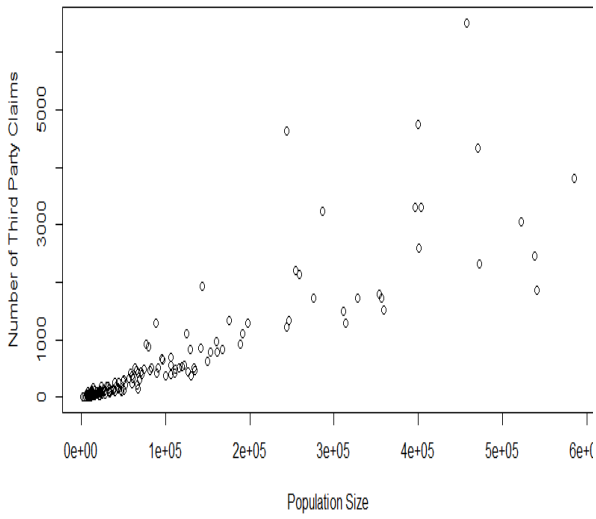


Figure 7. Pop vs. Third Party

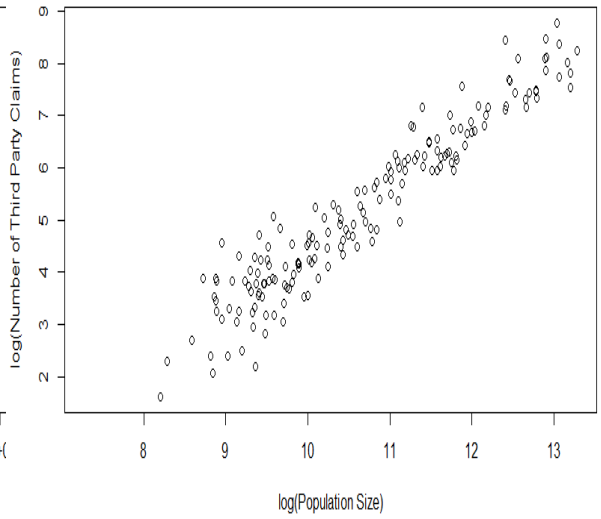


Figure 8. Pop vs. Third Party (log scale)

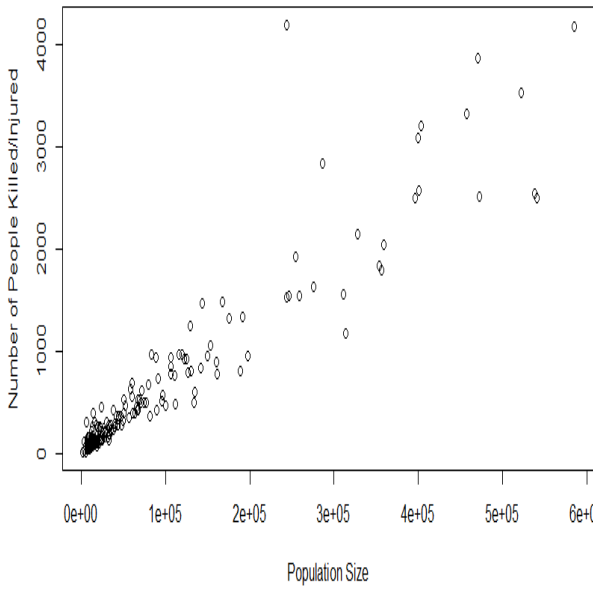


Figure 9. Pop vs. K/I

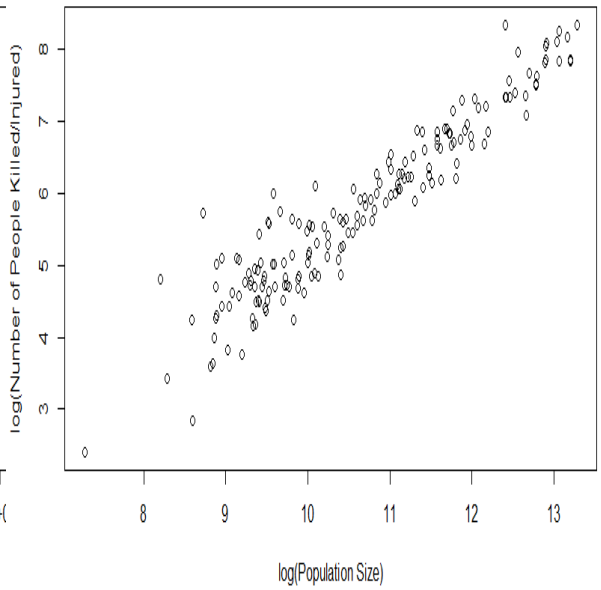


Figure 10. Pop vs. K/I (log scale)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

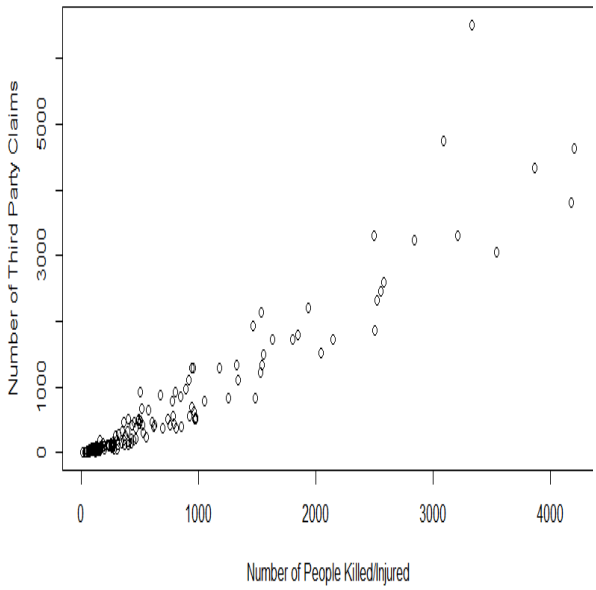


Figure 11. K/I vs. TPC

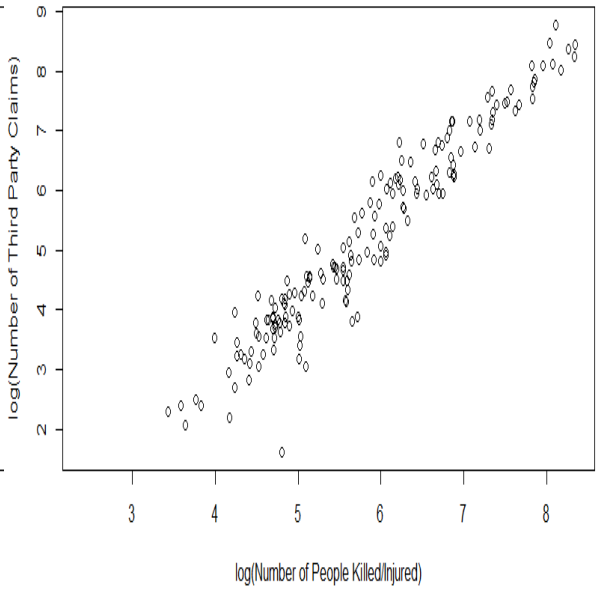
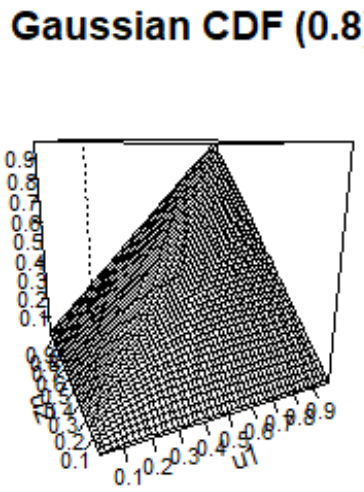


Figure 12. K/I vs. TPC (log scale)

Next, we provide plots of the associated c.d.fs and p.d.fs for the copulas discussed above.



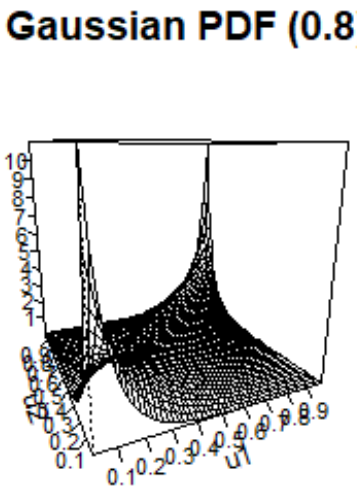


Figure 13. Gaussian (0.8) CDF and PDF

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

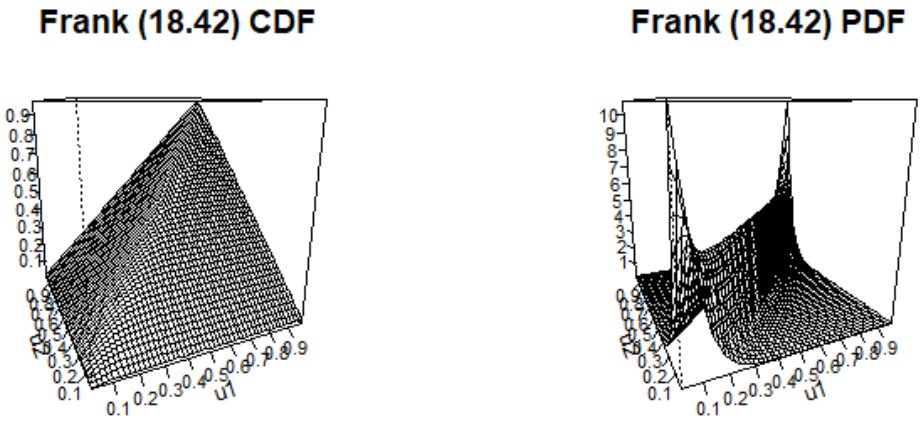


Figure 14. Frank CDF and PDF with $\alpha = 18.42$

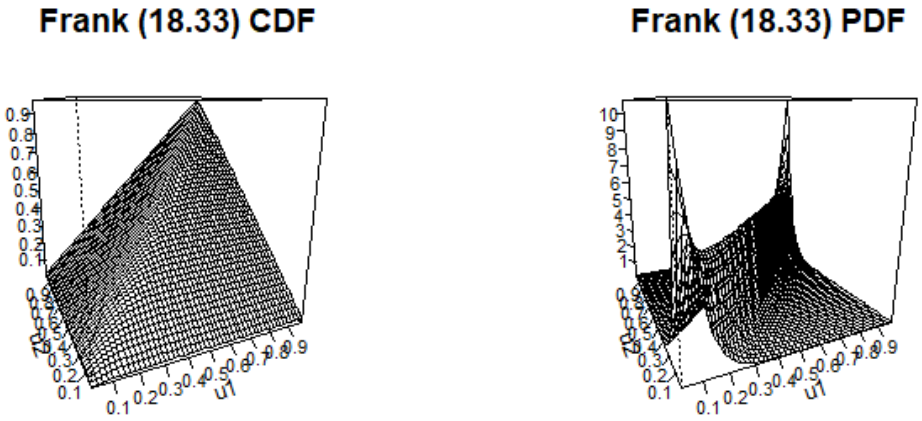


Figure 15. Frank CDF and PDF with $\alpha = 18.33$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

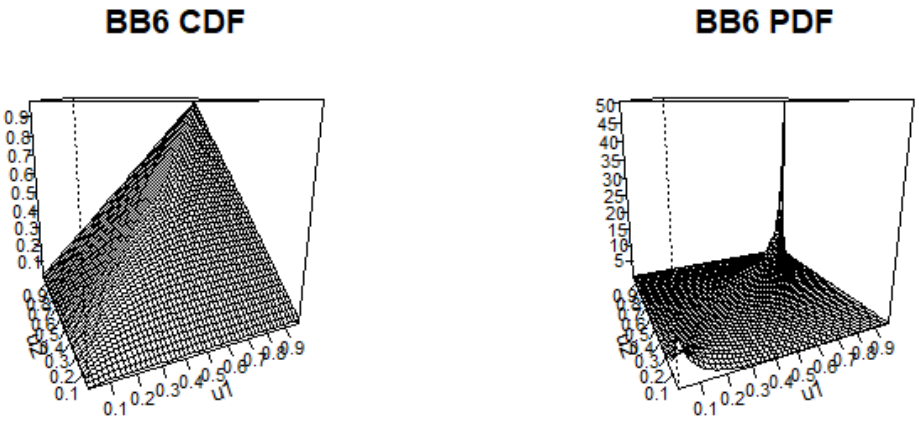


Figure 16. BB6 CDF and PDF with $\theta = 1.59$ and $\delta = 2.81$

4. Structural Properties

This section presents the analysis of certain structural properties of the copulas. Since the Gaussian, Tawn, and T copulas do not have a closed form, we exclude them from our current discussion. We begin our discussion with the Frank copula.

4.1. Frank Copula

The Frank Copula has the form:

$$C(u,v) = \log_\alpha \left[1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1}\right],$$

where $\alpha \geq 0$. We will begin our analysis of the Frank copula by calculating the Blomqvist Beta correlation coefficient using (11). Using our Australian insurance models whose best fitted copula was the Frank copula, we have the following direct calculations using $\alpha_1 = 18.42$ and $\alpha_2 = 18.33$:

$$\beta_1 = 4\log_{18.42}\left[1 + \frac{((18.42)^{1/2} - 1)^2}{17.42}\right] - 1 = 0.9348;$$
$$\beta_2 = 4\log_{18.33}\left[1 + \frac{((18.33)^{1/2} - 1)^2}{17.33}\right] - 1 = 0.9329.$$

Therefore, we see that both models are highly correlated in the center of their respective distributions. Next, we will determine the level of tail dependence in the general Frank copula. For upper tail dependence, from (9),

$$\lambda_U = \lim_{u\uparrow 1} \frac{1 - 2u + \log_\alpha\left[1 + \frac{(\alpha^u - 1)^2}{(\alpha - 1}\right]}{1 - u} \overset{H}{=} \lim_{u\uparrow 1} -\left(-2 + \left(\frac{1}{1 + \frac{(\alpha^u - 1)^2}{\alpha - 1}}\right)\left(\frac{2(\alpha^u - 1)(\alpha^u \ln \alpha)}{(\alpha - 1)\ln \alpha}\right)\right) = 0$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

Therefore, Frank's copula is not upper tail dependent. Next, we will determine if it is lower tail dependent using (10):

$$\lambda_L = \lim_{u\downarrow 0} \frac{\log_\alpha\left[1+\frac{(\alpha^u-1)^2}{\alpha-1}\right]}{u} \overset{H}{=} \lim_{u\downarrow 0} \frac{2(\alpha^u-1)(\alpha^u)(\ln\alpha)^2}{1+\frac{(\alpha^u-1)^2}{\alpha-1}} = 0$$

Hence the Frank Copula is also lower tail independent. The table below summarizes the dependence structures discussed above and displays the generator function of this particular copula:

Table 4. Dependence Structures of the Frank Copula

Generator Function	$\phi(t) = -\log\left(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right)$
Blomqvist Beta (General)	$\beta = 4\left(\log\left(1+\frac{(\sqrt{\alpha}-1)^2}{\alpha-1}\right)\right)-1$
Blomqvist Beta (AUS 1)	0.9348
Blomqvist Beta (AUS 2)	0.9329
Upper Tail Dependence	0
Lower Tail Dependence	0
Kendall's τ	$1+\frac{4(D_1(\alpha)-1)}{\alpha}$

Where $D_1 = \frac{1}{\alpha}\int_0^\alpha \frac{t}{e^t-1}dt$ is the Debye function of type 1.

4.2. BB6 (Joe-Gumbel) Copula

The BB6 copula has the following form:

$$C(u,v) = 1-\left(1-\exp(-[(-\log(1-\overline{u}^\theta))^\delta + (-\log(1-\overline{v}^\theta))^\delta]^{\frac{1}{\delta}})\right)^{\frac{1}{\theta}}, \quad u\geq 0, v\leq 1, \theta\geq 1, \delta\geq 1.$$

where $\overline{u} = 1-u$ and $\overline{v} = 1-v$. We will first determine this copula's Blomqvist Beta value based on the parameters of the Swedish Automobile insurance model with $\theta = 1.59$ and $\delta = 2.81$. We then have the following by applying (1):

$$\beta = 4\left(1-\left(1-\exp\left[-\left(2\left(-\log\left(1-\left(\frac{1}{2}\right)^{1.59}\right)\right)^{2.81}\right)^{\frac{1}{2.81}}\right]\right)^{\frac{1}{1.59}}\right)-1 = 0.7397$$

The lower tail and upper tail dependence coefficients can be calculated using the same methodology that we used for the Frank copula. For the upper tail dependence coefficient, we obtain the following:

$$\begin{aligned}\lambda_U &= \lim_{u\uparrow 1}\frac{1-2u+1-\left(1-\exp(-[2(-\log(1-\overline{u}^\theta))^\delta])\right)^{\frac{1}{\theta}}}{1-u}\\ &\overset{H}{=} \lim_{u\uparrow 1} 2-2^{\frac{1}{\delta}}(1-u)^{\theta-1}\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^\theta)\right](1-\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^\delta]\right)^{\frac{1}{\theta}-1} = 2-2^{\frac{1}{\theta\delta}}\end{aligned}$$

Similarly, for the lower tail dependence coefficient:

$$\begin{aligned}\lambda_L &= \lim_{u\downarrow 0}\frac{1-\left(1-\exp(-[2(-\log(1-\overline{u}^\theta))^\delta])\right)^{\frac{1}{\theta}}}{u}\\ &\overset{H}{=} \lim_{u\downarrow} 2^{\frac{1}{\delta}}(1-u)^{\theta-1}\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^\theta)\right](1-\exp\left[2^{\frac{1}{\delta}}\log(1-(1-u)^\delta]\right)^{\frac{1}{\theta}-1} = 0.\end{aligned}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2019 doi:10.20944/preprints201906.0235.v1

Table 5. Dependence Structures of the BB6 Copula

Generator Function	$\phi(t) = (-\log[1-(1-t)^{\theta}])^{\delta}$
Blomqvist Beta (General)	$\beta = 4\left(1-\left(1-e^{-[(-\log(1-\frac{1}{2}^{\theta}))^{\delta}+(-\log(1-\frac{1}{2}^{\theta}))^{\delta}]^{\frac{1}{\delta}}}\right)^{\frac{1}{\theta}}\right)$
Blomqvist Beta (Swedish Auto)	0.7397
Upper Tail Dependence(General)	$2-2^{\frac{1}{\delta\theta}}$
Upper Tail (Swedish Auto)	0.8321
Lower Tail Dependence	0
Kendall's τ	$1+\frac{4}{\delta\theta}\int_0^1(-\log(1-(1-t)^{\theta})(1-t)(1-(1-t)^{-\theta})dt$

5. Conclusion

In this article, we have utilized Vine Copulas with Frank and BB6 families of copula based on R-package VineCopula for fitting several well-known insurance data sets. In addition, we have also provided some structural properties of the most appropriate bivariate copulas including LTD and RTI property. The goodness of fit statistics are provided in terms of AIC and BIC values as well as the log-likelihood values. As a future research, we will be focusing on data sets from other domain (such as health care data), and also we will consider the fitting to a trivariate and in higher dimensions as well based on the Vine copula methodology. We will report our findings in a separate article.

Appendix

R package: Vine Copula Here, we provide a generic R-code based on the Vine Copula package which is used in the main body of the text for selecting the best possible bivariate copula on the four different insurance data set:

[language=R] install.packages("copula")

```
library(("copula")
m<-pobs(a)
n<-pobs(b)
install.packages("VineCopula")
library("VineCopula")
selectedCopula <- BiCopSelect(m, n, familyset = NA)
summary(selectedCopula)
```

Remark: In the above code a and b are the transformed (on a Log (to the base e) scale) variable values corresponding to two components of the associated bivariate data.

The best fitted bivariate copulas mentioned here do not possess a closed form of expression in terms of their density function (i.e., the p.d.f.). However, in order to get the p.d.f. of each of these copulas, one may use *R*. Next, we provide an example as to how one can simulate from the p.d.f. of a Survival BB1 copula with specific parameter choices in *R*.

[language=R] Simulate from a bivariate rotated BB1 copula (180 degrees; "survival BB1")

```
install.packages("VineCopula")
library("VineCopula")
SBB1<- BiCop(family = 17, par =0.63, par2 =1.09)
sim<- BiCopSim(1000, SBB1)
```

To evaluate the density of the bivariate rotated BB1 copula

```
x1<- simdata[,1]
x2 <- simdata[,2]
BiCopPDF(u1, u2, SBB1)
```

References

1. Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics,* 44,182–198.
2. Balakrishnan, N., and Lai, C.D. (2009). Continuous Bivariate Distributions, Second edition. Springer, New York.
3. Bedford, T., Cooke, R.M. (2001). Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Annals of Mathematics and Artificial Intelligence,* 32, 245-268.
4. Bedford, T., Cooke, R.M. (2002). Vines – A New Graphical Model for Dependent Random Variables. *The Annals of Statistics,* 30, 1031–1068.
5. Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statististics,* 21, 593-600.
6. Brechmann, E.C., Czado, C. (2013). Risk Management with High-Dimensional Vine Copulas: An Analysis of the Euro Stoxx 50. *Statistics & Risk Modeling*, 30, DOI: 10.1524/strm.2013.2002.
7. Brechmann, E.C., Czado, C., and Aas, K. (2012). Truncated Regular Vines in High Dimensions with Applications to Financial Data. *Canadian Journal of Statistics,* 40, 68-85.
8. Chen, X., Fan, Y. (2005). Pseudo-likelihood ratio tests for model selection in semiparametric multivariate copula models. *The Canadian Journal of Statistics,* 33, 389-414.
9. Chollete, L., Heinen, A., Valdesogo, A. (2009). Modeling International Financial Returns with a Multivariate Regime Switching Copula. *Journal of Financial Econometrics,* 7, 437-480.
10. Cook, R. D., and Johnson, M.E. (1981). A family of distributions for modeling non-elliptically symmetric multivariate data. *Journal of the Royal Statististical Society, Series B,* 43, 210-218.
11. Cook, R. D., and Johnson, M.E. (1986). Generalized Burr-Pareto-logistic distributions with applications to a uranium exploration data set. *Technometrics,* 28, 123-131.
12. Czado, C. (2010). Pair-Copula Constructions of Multivariate Copulas. In P Jaworski, F Duante, W Hardle, T Rychlik (eds.), *Copula Theory and Its Applications.* Springer-Verlag, Berlin
13. Czado, C., Schepsmeier, U., Min, A. (2012). Maximum Likelihood Estimation of Mixed C-Vines with Application to Exchange Rates. *Statistical Modeling,* 12, 229-255.
14. Dall'Aglio, G. (1991). Frechet classes: the beginnings. Advances in probability distributions with given marginals (Rome, 1990). *Math. Appl.,Kluwer Acad. Publ., Dordrecht* 67, 1-12.
15. de Melo Mendes, B.V., Mendes Semeraro, M., C'amara Leal, R.P. (2010). Pair-Copulas Modeling in Finance. *Financial Markets and Portfolio Management,* 24, 193-213.
16. Embrechts, P., KlAsuppelberg, C., and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance.* Springer, Berlin.
17. Frees, E. W., and Valdez, E.A. (1998). Understanding relationships using copulas. *North Am. Act. J.* 2, 1-25.
18. Genest, C., Quessy, J.R., and Remillard, B. (2006). Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation. *Scandinavian Journal of Statistics,* DOI:10.1111/j.1467-9469.2006.00470.
19. Ghosh, I., and Ray, S. (2016). Some alternative bivariate Kumaraswamy type distributions via copula with application in risk management. *Journal of Statistical Theory and Practice,*10, 693-706.
20. Joe, H. (1996). Families of m-Variate Distributions with Given Margins and $m(m-1)/2$ Bivariate Dependence Parameters. In Ruschendorf, B Schweizer, MD Taylor (eds.), *Distributions with Fixed Marginals and Related Topics,* pp. 120-141. Institute of Mathematical Statistics, Hayward, CA.
21. Joe, H. (1997). *Multivariate Models and Dependence Concepts,* Chapman & Hall, New York.
22. Joe, H., Li, H., Nikoloulopoulos, A.K. (2010). Tail Dependence Functions and Vine Copulas. *Journal of Multivariate Analysis,* 101, 252-270.
23. Haug, S., Kluppelberg, C., and Kuhn, G. (2011). Statistical models and methods for dependence in insurance data. *Journal of the Korean Statistical Society,* 40, 125-139.
24. Heinen, A., Valdesogo, A. (2009). *Asymmetric CAPM Dependence for Large Dimensions: The Canonical Vine Autoregressive Model.* CORE discussion papers 2009069, Universit'e catholique de Louvain, Center for Operations Research and Econometrics (CORE).
25. Hofmann, M., Czado, C. (2010). Assessing the VaR of a Portfolio Using D-Vine Copula Based Multivariate GARCH Models. *Submitted for publication.*

26. Klugman, S. A., and R. Parsa. (1999). Fitting bivariate loss distributions with copulas. *Insur. Math. Econ.* 24, 139-148.
27. Kurowicka, D., Cooke, R.M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modeling.* John Wiley & Sons, Chichester.
28. Kurowicka, D., Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook.* World Scientific Publishing Co., Singapore.
29. Min, A., Czado, C. (2010). Bayesian Inference for Multivariate Copulas Using Pair-Copula Constructions. *Journal of Financial Econometrics,* 8, 511-546.
30. Min, A., Czado, C. (2011). Bayesian Model Selection for Multivariate Copulas Using Pair- Copula Constructions. *Canadian Journal of Statistics,* 39, 239-258.
31. Nikoloulopoulos, A.K., Joe, H., and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data.*Computational Statistics & Data Analysis,* 56, 3659-3673.
32. Nelsen, R.B. (1999). *An Introduction to Copulas,* first edition, Springer-Verlag, New York.
33. Nelsen, R.B. (2006). *An Introduction to Copulas,* second edition, Springer-Verlag, New York.
34. Schirmacher, D., and Schirmacher, E. (2008). *Multivariate Dependence Modeling Using Pair- Copulas.* Technical report, Society of Acturaries: 2008 Enterprise Risk Management Symposium, April 14-16, Chicago.
35. Schweizer, B. and Sklar, A. (1983). *Probabilistic metric spaces.* North-Holland Publishing Company, New York.
36. Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *(French) Publ. Inst. Statist. Univ. Paris,* 8, 229-231.
37. Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika (Prague),* 9, 449-460.
38. Smith, M., Min, A., Almeida, C., and Czado, C. (2010). Modeling Longitudinal Data Using a Pair-Copula Decomposition of Serial Dependence. *Journal of the American Statistical Association,* 105, 1467–1479.