Life Expectancy Estimate With Bivariate Weibull Distribution Using Archimedean Copula

Eun-Joo Lee elee@millikin.edu
Department of Mathematics
Millikin University
Decatur, IL 62522

Chang-Hyun Kim maraychk@gmail.com
Illinois Natural History Survey
University of Illinois at Urbana-Champaign
Champaign, IL 61820

Seung-Hwan Lee slee2@iwu.edu
Department of Mathematics and Computer Science
Illinois Wesleyan University
Bloomington, IL 61701

Abstract

Archimedean copulas are used to construct bivariate Weibull distributions. Co-movement structures of variables are analyzed through the copulas, where the tail dependence between the variables is explored with more flexibility. Based on the distance between the copula distribution and its empirical version, a copula that may best fit data is selected. With extra computing costs, the adequacy of the copula chosen is then assessed. When multiple myeloma data are considered, it is found that relationship between survival time of a patient and the hemoglobin level is well described by the Clayton copula. The bivariate Weibull distribution constructed by the copula is used to estimate value at risk from which we investigate the anticipated longest life expectancy of a patient with the disease over the treatment period.

Keywords: Archimedean Copula, Dependence, Weibull Distribution, Value at Risk.

1. INTRODUCTION

Copulas are a useful tool used to model a joint distribution function of variables of interest. In particular, copulas have gained their importance as simple functions to describe the dependence structure of random variables in the joint distribution. As a model for the dependence structure, copulas have several advantages over other dependence measures such as the correlation coefficient (Sklar [28], Genest and Rivest [12], Nelson [26]). For example, using copulas, modeling both linear and non-linear dependencies of variables is possible, and the degree of dependence in the tail of the underlying distribution can be described (Embrechts et al. [7]). Many authors have studied the use of copula in applications, including risk management (Freez and Valdez [9]) and survival analysis (Zheng and Klein [30], Rivest and Wells [27]). In this work, we construct bivariate Weibull distributions using Archimedean copulas that reflect on the asymmetric dependence structure. These copulas are the Gumbel, Clayton, Frank and Independence copulas, each having different characteristics of tail dependence.

Copulas have varying amounts of tail dependence depending on the choice of copulas. Therefore, an important issue in using copulas is the choice of appropriate copulas. Poorly chosen copulas may lead to undesired results about the actual relationship between variables. The copula selection issue has been studied by many authors, including Melchiori [25], Durrelman et al. [6], Kumar and Shoukri [19], Frees and Valdez [9] and Genest and Rivest [12]. Similar to the procedures they have developed, we discuss the copula selection procedures based on the distance of the copula distribution and its empirical version. With extra computing costs, we further examine the goodness of fit of the copula selected. The procedures are based on a process over the domain of the generator for Archimedean copulas. Under the null hypothesis of the no model misspecification, the distributions of the process from the distance measure can be easily approximated by the simulation technique. As a numerical measure for the assessment of the model adequacy, we consider the supremum of the process from which the empirical p -values are obtained.

Multiple myeloma is a progressive and invariably fatal disease caused by the accumulation of abnormal plasma cells in the bone marrow. The prognosis of the disease is often unpredictable and overall survival is ranged from a few months to more than 10 years (Kyle and Rajkumar [20]). Traditionally, multiple myeloma has been staged by the method developed by Durie and Salmon [5], although a newer staging method has been developed recently (Greipp et al. [13]). In the staging method by Durie and Salmon [5], it has been known that the level of hemoglobin (denoted by HB hereafter) in the blood of a multiple myeloma patient is strongly associated with the tumor mass and thus is a strong indicator of the disease progress (Durie and Salmon [5], Kyle and Rajkumar [20]). The objective of this paper is to demonstrate the benefits of using copulas to model dependencies in multiple myeloma data with a particular focus on potential survival time of a patient over the treatment period. This was carried out at the Medical Centre at the University of West Virginia. To simplify our discussion, the complete data points of survival time, in months, of male patients with the disease (denote by ST hereafter), i.e., the time from diagnosis until death from multiple myeloma, and the corresponding level of HB are considered in this paper, where the sample size is 22. See Krall et al. [18] and Collect [2] for details about the data. The effect of HB on the survival times of the patients is explored using the bivariate Weibull distribution constructed by copulas, where a measure of linear dependence is not so informative, as will be seen in Figure 2 and described in Section 4.2. We incorporate the copulas into the calculation of value at risk for the survival time. The value at risk is a risk measurement technique often used in the area of financial risk management (Jorion [17]). We use this method as a tool to estimate the anticipated maximum life span, i.e. maximum extension possible for a life, with reference to the level of hemoglobin that influences the survival time.

The layout of this paper is as follows. Section 2 presents Archimedean copula functions used in this work. Section 3 discusses the association parameter and the dependence measure of the copula functions. Section 4 constructs bivariate Weibull distributions using copula, checks the adequacy of the copula selected, and calculates value at risk associated with survival time. Concluding remarks are presented in Section 5.

2. COPULA MODEL

2.1 Copula Function

Estimating a multivariate distribution with correlations is not an easy process. A copula is a useful tool that accommodates this problem. It joins a multivariate distribution function to univariate marginal distribution functions, so a copula function is a multivariate distribution function. Specifically, a copula function, denoted by C, is a multivariate distribution function with uniform marginal distribution functions, $F_1, F_2, ..., F_p$, on the interval [0, 1], i.e., if for $x_1, x_2, ..., x_p$, $F(x_1, x_2, ..., x_p)$ is a multivariate probability distribution with marginals $F_1(x_1), F_2(x_2), ..., F_p(x_p)$, then $F(x_1, x_2, ..., x_p)$ can be written as

$$F(x_1, x_2, ..., x_p) = C(F_1(x_1), F_2(x_2), ..., F_p(x_p)).$$

For a bivariate case, the copula form is the easiest way to express and generate the joint distribution (Venter [29]). In this work, we primarily look at this bivariate copula. In the bivariate case, a copula is a function $C:[0,1]^2 \to [0,1]$ such that $C(u,0) = C(0,u) = 0$ for all u in $[0,1]$, $C(v,1) = C(1,v) = v$ for all v in $[0,1]$ and $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$

for all $0 \le u_1 \le u_2 \le 1$ and $0 \le v_1 \le v_2 \le 1$. From this, when $F_1(x_1) = u$ and $F_2(x_2) = v$, a copula function $C(F_1(x_1), F_2(x_2))$ is a proper bivariate distribution function. Conversely, any bivariate distribution function $F(x_1, x_2)$ with continuous marginal distribution functions F_1 and F_2 can be uniquely expressed by a copula function

$$C(u,v) = F(F_1^{-1}(u), F_2^{-1}(v)).$$

The following theorem summarizes the above results (Sklar [28]).

Theorem (Sklar's Theorem) Let F be a bivariate joint distribution function of continuous random variables X and Y with corresponding marginal distribution functions F_1 and F_2. There exists a copula C (i.e., a bivariate distribution function on $[0,1]^2$ with uniform marginal distribution functions) such that, for $-\infty < x_1 < \infty, -\infty < x_2 < \infty$,

$$F(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2) = C(F_1(x_1), F_2(x_2)). \qquad (1)$$

Note that Sklar's theorem simply implies that $C(u,v) = P(U \le u, V \le v)$ for uniform random variables U and V over $[0,1]$. We close this section by describing meaningful bounds for copula.

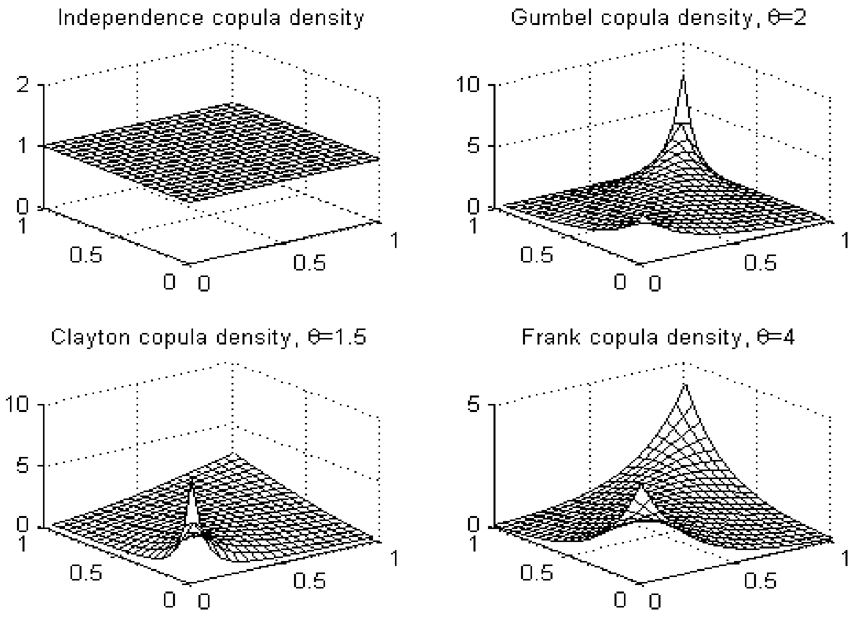
Theorem (Frechet-Hoeffding Bounds) For every copula C and every (u,v) in $[0,1]^2$,

$$\max(u + v - 1, 0) \le C(u,v) \le \min(u,v).$$

Note that by Sklar's theorem $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$, where $C(u,v) = F(F_1^{-1}(u), F_2^{-1}(v))$ and $u = F_1(x_1), v = F_2(x_2)$. Thus,

$$\max(F_1(x_1) + F_2(x_2) - 1, 0) \le F(x_1, x_2) \le \min(F_1(x_1), F_2(x_2)).$$

2.2 Archimedean Copula


Independence copula density

Gumbel copula density, θ=2

Clayton copula density, θ=1.5

Frank copula density, θ=4

FIGURE1: Independence, Gumbel, Clayton, Frank copula density plots

The Archimedean copula is a convenient method to model a bivariate distribution due to its simple form and a variety of dependence structures. The use of Laplace transformations leads to the construction of Archimedean copulas. Specifically, let φ be a continuous monotonically decreasing function from $[0,1]$ to $[0,\infty)$ such that $\varphi(1)=0$ and $\varphi^{''}(x)>0$. Define the pseudo inverse of φ as follows: $\varphi^{[-1]}(x)=\varphi^{-1}(x)$ for $0\le x\le \varphi(x)$ and zero for $\varphi(0)\le x\le\infty$. Note that if $\varphi(0)=\infty$, then $\varphi^{[-1]}=\varphi^{-1}$. Then, for real numbers, u and v, an Archimedean copula, C, of bivariate random variables U and V is given by

$$C(u,v)=\varphi^{[-1]}(\varphi(u)+\varphi(v)), \qquad (2)$$

where $\varphi:[0,1]\to[0,\infty)$, and this φ is often called a generator of copula.

Archimedean copulas involve a tail dependence parameter, also referred as the association parameter. This describes the amount of dependence in the upper tail or lower tail of a multivariate distribution and can be used to analyze the dependence among extreme values. The generator φ contains all of the information about the dependence structure of the multivariate distribution of random variables in terms of the parameter of association. The association parameter is denoted by θ throughout this paper.

Based on the level of the tail dependence structure, we consider four families of Archimedean copulas in this paper. They are the Gumbel copula (Gumbel [14], Hougaard [15]), the Clayton copula (Clayton [1]) which is also referred to as Cook and Johnson's copula (Cook and Johnson [3]), the Frank copula (Frank [8]) and Independence (or Product) copula. Each family of the copulas is generated by the formula (2) through the generator φ . Specifically, Gumbel's copula is

$$C(u,v;\theta)=\exp\{-[(-\log u)^{\theta}+(-\log v)^{\theta}]^{1/\theta}\},$$

and the corresponding generator is $\varphi(t)=(-\log t)^{\theta}$ for $\theta\ge 1$. As a special case, the parameter $\theta=1$ implies independence between the distributions. With $\theta\to\infty$, the Gumbel copula attains the Frechet-Hoeffding lower bound, so the distribution is characterized by extreme values. This implies higher dependence in the upper tail. Clayton's copula is

$$C(u,v;\theta)=(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}.$$

The Clayton copula generator is $\varphi(t)=\dfrac{t^{-\theta}-1}{\theta}$ for $\theta>0$. With $\theta\to\infty$, the Clayton copula attains the Frechet-Hoeffding upper bound, so higher dependence in the lower tail. As $\theta\to 0$, the Clayton copula implies independence between the distributions. The Frank copula is

$$C(u,v;\theta)=-\frac{1}{\theta}\log[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}],$$

and its generator is $\varphi(t)=-\log[\dfrac{e^{-\theta t}-1}{e^{-\theta}-1}]$, $\theta\ne 0$. With $\theta\to\infty$, the Frank copula attains the Frechet-Hoeffding upper bound, and with $\theta\to-\infty$, it attains the Frechet-Hoeffding lower bound. When $\theta>0$ ($\theta<0$), it implies positive (negative) dependence between the distributions. When $\theta\to 0$, the copula implies independence between the distributions. Finally, the independence copula is

$$C(u,v)=u\cdot v,$$

with the generator $\varphi(t)=-\log t$. Note that there is no association parameter in the independence copula.

Tail dependence of copulas can be illustrated by their density function. The bivariate distribution function is defined in (1), and the corresponding density function is obtained by differentiation,

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

where c is the density of C, and f_1 and f_2 are the marginals. Written this way, it is also possible to define Archimedean copulas in the multivariate case. See McNeil et al. [24] for details. Figure 1 displays the copula densities. As depicted from this figure, each copula has varying degrees of dependence according to values of θ. Note in this figure that the independence copula has unit everywhere. The estimation of parameters of the copulas is discussed in Section 3.1.

3. ASSOCIATION AND DEPENDENCE

3.1 Measuring Association

Two commonly used measures of association would be Spearman's ρ and Kendall's τ. They are based on the rank of data, so they have the invariance property under monotonic transformations. For Archimedean copulas, Kendall's τ has the copula representation, and so it captures perfect dependence. On the contrary, there seems to be no simple expression for Spearman's ρ. Kendall's τ is given by (Nelson [26], Joe [16], Genest and Mackay [10, 11])

$$\tau=4\int_0^1\int_0^1 C(u,v)dC(u,v)-1. \tag{3}$$

From the expression in (3), Kendall's τ is calculated by a copula that contains the association parameter. Conversely, the association parameter can be measured by Kendall's τ obtained from data. For bivariate Archimedean copulas, where the two random variables are absolutely continuous, Kendall's τ can be readily calculated via the following identity (Genest and MacKay [10, 11])

$$\tau=1+4\int_0^1\frac{\varphi(t)}{\varphi'(t)}dt\ . \tag{4}$$

The copula generator contains the association parameter, and from (4) the generator can be expressed through Kendall's τ. Therefore, the association parameter, θ, is measured by solving the identity in (4). For example, it can be shown that for the Clayton copula,

$$1+4\int_0^1\frac{\varphi(t)}{\varphi'(t)}dt=\frac{\theta}{\theta+2},$$

and this yields $\tau=\dfrac{\theta}{\theta+2}$. Similar algebra leads to $\tau=\dfrac{\theta-1}{\theta}$ for the Gumbel copula. Unlike these two copulas, the Frank copula doesn't have a closed form of τ that can be directly expressed by θ. It is necessary to use numerical methods to solve the following identity

$$\tau=1-\frac{4}{\theta}(1-\frac{1}{\theta}\int_0^\theta\frac{t}{e^t-1}dt).$$

In this work, the random search method is utilized to estimate the Frank copula parameter, among other numerical methods.

The dependence structure of HB and ST is displayed in the scatter plot (Figure 2), where Kendall's τ is 0.2208. This and the procedures above result in θ= 1.2834, 0.5667 and 2.07 for the Gumbel, Clayton and Frank copulas, respectively.

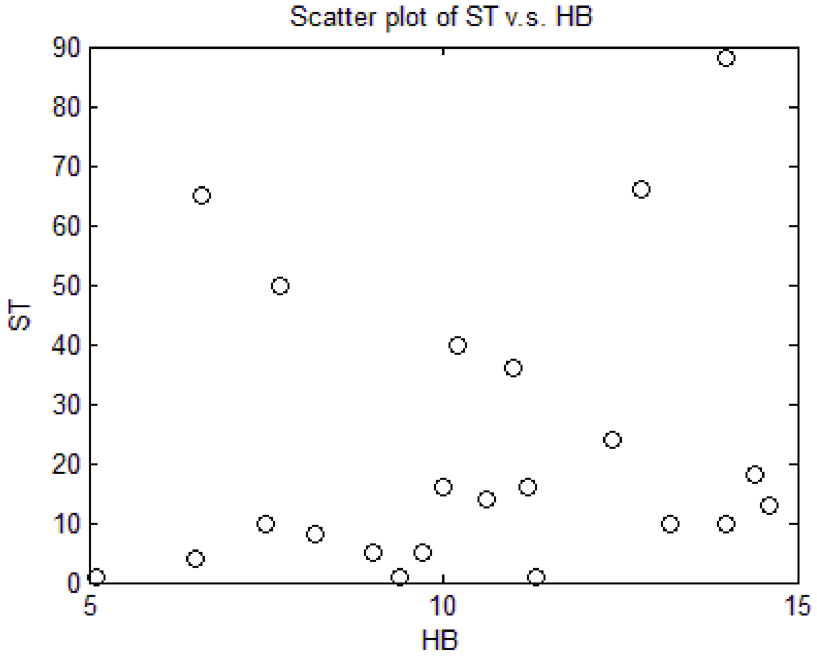


FIGURE 2: Scatter plot of HB vs ST

3.2 Dependence

Tail dependence deals with the degree of dependence in the tails of a bivariate distribution, so it describes the dependence structure of extreme events. For example, Figure 1 in Section 2.2 displays that different copulas show different behaviors in their tails. This implies that tail dependence may vary depending on the choice of copulas.

Let X_1 and X_2 be random variables with continuous distribution functions $F_1(x_1)$ and $F_2(x_2)$. The upper- and lower-tail dependence coefficients are defined as the limit of conditional probability, respectively,

$$\lambda_U = \lim_{u \to 1-} P(Y \geq F_2^{-1}(u) \mid X \geq F_1^{-1}(u)),$$

$$\lambda_L = \lim_{u \to 0+} P(Y \leq F_2^{-1}(u) \mid X \leq F_1^{-1}(u)),$$

for u in (0,1). If the value of the upper (lower) tail dependence coefficient is positive, then X_1 and X_2 have structure dependent at the upper (lower) tail. In contrast, zero tail dependence implies asymptotic independence. The tail dependences can be also expressed through copula, showing the fact that the tail dependence is a copula property,

$$\lambda_U = \lim_{u \to 1-} \frac{1-2u+C(u,u)}{1-u}, \quad \lambda_L = \lim_{u \to 0+} \frac{C(u,u)}{u}.$$

From this, the Gumbel, Clayton, Frank and independence copulas have $[0, 2-2^{\theta}]$, $[2^{-1/\theta}, 0]$, $[0,0]$ and $[0,0]$, respectively, where [a, b] = [lower, upper] tail dependence coefficients. By plugging in the values of θ found in Section 3.1, theoretical [lower, upper] tail dependence coefficients for the Gumbel, Clayton, Frank and independence copulas are [0,0.2838], [0.2943,0], [0,0] and [0,0], respectively. This indicates that the Gumbel copula has the upper tail dependence but does not have the lower tail dependence, the Clayton copula has the lower tail dependence but does not have the upper tail dependence, while the Frank and the independence copulas have neither. Numerical computations of the tail dependence using the limit formulas above are reported in Table 1. It shows the same phenomena as in the theoretical analysis. For example, as u tends to 1 through values less than 1, λ_U for Gumbel tends to 0.2838.

	Gumbel	Clayton	Frank	Indep.		Gumbel	Clayton	Frank	Indep.
$u \to 0+$	λ_L	λ_L	λ_L	λ_L .	$u \to 1-$	λ_U	λ_U	λ_U	λ_U
.10	0.1922	0.3806	0.1973	0.1000	.90	0.3459	0.1482	0.1973	0.1000
.05	0.1170	0.3486	0.1075	0.0500	.95	0.3147	0.0762	0.1075	0.0500
.005	0.0225	0.3076	0.0117	0.0050	.995	0.2869	0.0078	0.0117	0.0050
.001	0.0071	0.2995	0.0024	0.0010	.999	0.2844	0.0016	0.0024	0.0010
.00001	0.0003	0.2947	0.0000	0.0000	.99999	0.2838	0.0000	0.0000	0.0000

TABLE 1: Tail dependence coefficient for the copulas associated with data

4. THE PROCEDURES

4.1 Bivariate Weibull Distribution

In the parametric analysis of survival analysis, one of the commonly used models is the two-parameter Weibull distribution. We construct the bivariate Weibull distributions based on the four copula functions stated in Section 2.2. Specifically, given two marginal Weibull distributions

$$F_i(x_i) = 1 - e^{-(x_i/\beta_i)^{\alpha_i}}, i = 1,2,$$

it is possible to construct a bivariate distribution $F(x_1, x_2)$ such that $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. For example, choosing the Gumbel copula gives a bivariate Weibull distribution given by

$$F(x_1, x_2) = \exp\{-[(-\log F_1(x_1))^{\theta} + (-\log F_2(x_2))^{\theta}]^{1/\theta}\},$$

the Clayton copula yields a bivariate Weibull distribution given by

$$F(x_1, x_2) = (F_1(x_1)^{-\theta} + F_2(x_2)^{-\theta} - 1)^{-1/\theta},$$

the Frank copula leads to a bivariate Weibull distribution given by

$$F(x_1, x_2) = -\frac{1}{\theta}\log[1 + \frac{(e^{-\theta F_1(x_1)} - 1)(e^{-\theta F_2(x_2)} - 1)}{e^{-\theta} - 1}],$$

and the independence copula produces a bivariate Weibull distribution given by

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

For the multiple myeloma data, where X_1 and X_2 respectively represent HB and ST, it is found that Weibull distributions can be fit to them with α_1 = 11.307, β_1=3.859, and α_2=20.175, β_2=0.839. Associated with these, we use the four bivariate distribution functions above as the underlying distribution of returns to compute value at risk stated in Section 4.4.

4.2 Simulation

From the scatter plot of ST and HB in Figure 2 in Section 3.1, Pearson's correlation coefficient is not sufficiently informative on the dependence structure. It is problematic to identify the dependence (co-movement) of the variables at the tails. Computation of the linear correlation coefficient of ST and HB yields 0.1852. The corresponding $p-$value for testing the hypothesis of no correlation against the alternative that there is a non-zero correlation is 0.4094. This value does not show pronounced evidence that the two variables are linearly dependent. So, the dependence structure of the variables is modeled via copula.

In Section 4.1, we created four bivariate Weibull distributions based on Gumbel, Clayton, Frank and the independence copulas. The first three copulas are respectively parameterized by θ=1.2834, 0.5667 and 2.07 as discovered in Section 3.1. Figure 3 shows 250 simulated values from the four bivariate Weibull distributions that use the Gumbel, Clayton, Frank and independence copulas. In this figure, it seems that the positive dependence is somewhat observed. A high level of hemoglobin (HB) tends to influence the survival time (ST) of patients in the Gumbel copula, indicating high tail dependence. Positive dependence between the variables is also observed in the Frank copula. However, it seems that there is no dominant copula among the Archimedean copulas considered here. The independence copula provides no distinct patterns due to the assumption of independence among the variables.

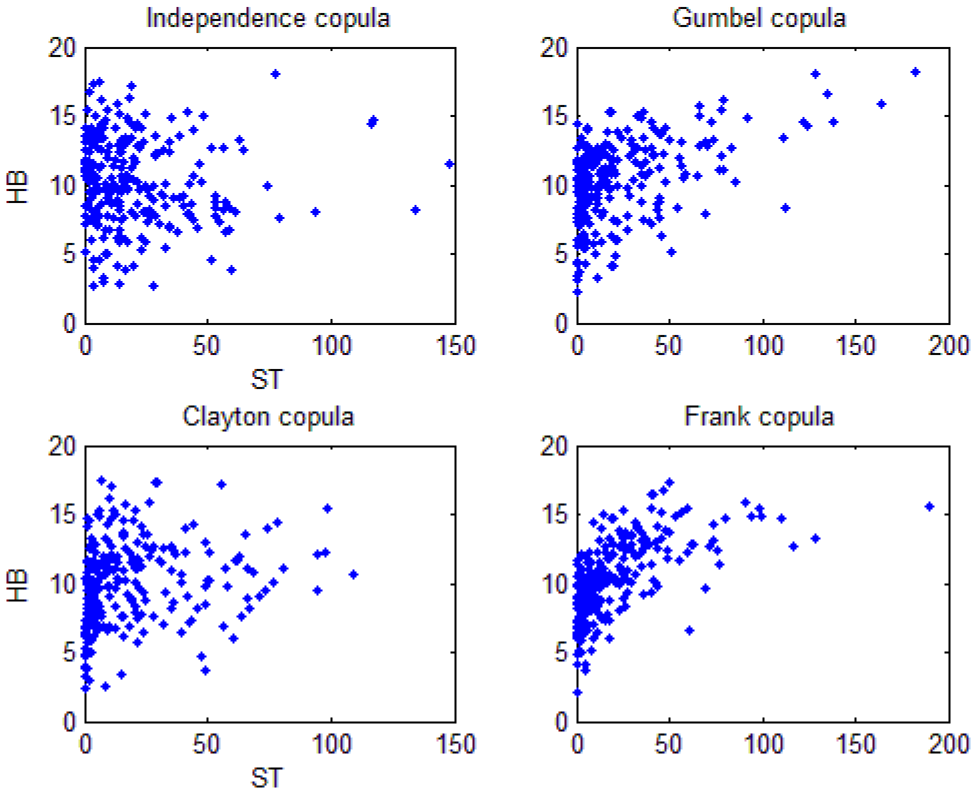


FIGURE 3: Indpendence, Gumbel, Clayton and Frank copulas, 1000 simulated samples

4.3 Copula Adequacy

A copula is the dependence structure of the data distribution. Since it has varying amounts of tail dependence depending on the choice of copulas, properly chosen copulas should be used in application. The adequacy of the copula selected also needs to be checked. These issues are discussed in this section. The procedures are based on the distance of the copula distribution and its empirical version.

Define a pseudo random variableT_i , for $i, j = 1,...,n$,

$$T_i = \{\text{the number of } (X_{1,i}, X_{2,j}), j = 1,...,n : X_{1,j} < X_{1,i}, X_{2,j} < X_{2,i}\}/(n-1) .$$

Further, define $K(t) = P(T_i \le t)$ for t in [0,1]. Genest and Rivest [12] showed that the distribution of $C(u,v)$ is

$$K(t) = t - \frac{\varphi(t)}{\varphi'(t)}.$$

By plugging in, the following $K(t)$'s for the copulas are obtained: the Gumbel copula has $K(t) = t - \frac{t \log t}{\theta}$, the Clayton copula takes $K(t) = t - \frac{t^{1+\theta} - t}{\theta}$, the Frank copula gives $K(t) = t - \frac{e^{t\theta} - t}{\theta} \log \frac{e^{-t\theta} - 1}{e^{-\theta} - 1}$, and the independence copula yields $K(t) = t - t\log t$. Now, define an empirical distribution of $K(t)$,

$$\hat{K}(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \le t),$$

where I is the indicator function. Then, we select the copula that best fits the data for which the distance of $K(t)$ and its estimate $\hat{K}(t)$ is minimized. Specifically, as usual in the literature, the best copula is selected as the one which minimizes the Kolmogorov-Smirnov type distance defined as

$$D(K, \hat{K}) = \int_0^1 \{K(t) - \hat{K}(t)\}^2 d\hat{K}(t). \tag{5}$$

For the Gumbel, Clayton, Frank and independence copulas associated with data, computation of $D(K, \hat{K})$ yields 0.0063, 0.0047, 0.0052 and 0.0162, respectively. This implies that the Clayton copula may be the best model for the data set considered. It appears that the independence copula is unlikely to be appropriate in this study.

We now check the validity of the copula chosen. The procedures are based on a process, derived from the distance measure in (5), over the domain of the copula generator. Define the process in t,

$$D(t) = \int_0^t \{K(t) - \hat{K}(t)\}^2 d\hat{K}(t)$$

for $0 < t \le 1$. Similar to Lin et al. [23], Lee and Yang [22] and Lee et al. [21], with the parameter of association, we generate data from the copula through simulation. From the simulated data, we obtain an estimate of the parameter, $\hat{\theta}$. Denote the resulting distribution of the copula by $K^*(t)$. The process associated with this and the parameter estimate is then given by

$$D^*(t, \hat{\theta}) = \int_0^t \{K^*(t) - \hat{K}(t)\}^2 d\hat{K}(t)$$

for $0 < t \le 1$, and this simulated process is an approximation to the observed process $D(t)$.

A large number of samples can be generated repeatedly from this simulated process. Since the null distribution of the copula is approximated by $D^*(t, \hat{\theta})$, there will be no distinguished behavior of $D(t)$ comparing to a large number of realizations produced from $D^*(t, \hat{\theta})$, if the copula fits the data. Under the null hypothesis that the copula model is valid, the process $D^*(t, \hat{\theta})$ will randomly fluctuate above, near zero. So, as a numerical measure for the assessment of the model adequacy, we consider the supremum of the process $D(t)$ over (0,1], $S = \sup_{0<t\le 1} D(t)$. An unusually large value of S would indicate that the copula is not valid. Let $S^* = \sup_{0<t\le 1} D^*(t, \hat{\theta})$. Then, the distribution of S is approximated by the conditional distribution of

S^* given the data. This implies that the p-value $P(S \geq s)$ can be approximated by $P(S^* \geq s)$, and $P(S^* \geq s)$ is estimated through the simulation technique. For the data considered here, using these procedures associated with the Clayton copula, the estimated p-value is 0.4945, which means the Clayton copula is appropriate. This estimated p-value is based on 1000 realizations of the simulated process as suggested by Lin et al. [23].

Using the same procedures, we also found that the results for Gumbel and Frank copulas are not significant (not shown). Therefore there is not sufficient evidence to reject the Gumbel, Clayton and Frank copulas. Thus all three classes of models may be applicable, although comparisons of the results from the individual models suggest that the Clayton copula may fit the data better. More data may be required to discriminate adequately between the three copulas.

4.4 Life Expectancy Estimate

Copula	VaR (90%)	ES (90%)	VaR (95%)	ES(95%)
Gumbel	54.6223	84.9359	74.8936	106.4190
Frank	54.5661	84.6025	74.7020	105.8340
Clayton	54.4268	84.3032	74.2875	105.5011
Independence	54.1984	83.6832	74.0877	104.4382

TABLE 2: Time estimates (in months) of VaR and ES

In this section, we employ the copulas to calculate value at risk. The value at risk (VaR) is a risk measurement technique often used in the area of Financial Risk Management (Jorion [17]). Consider a linear combination of X_1 and X_2, $Z = w_1X_1 + w_2X_2$, where X_1 and X_2 represent the same type of data, and w_1 and w_2 are the weights taken over the real number (R), for each variable, with distribution function F_Z. Let z be a realization of Z, and R be the set of real numbers. The Value at Risk of Z at probability level α is then defined as

$$\text{VaR}_{\alpha}(Z) = F_Z^{-1}(\alpha) = \inf\{z \in R \mid F(z) \geq \alpha\}.$$

Value at Risk is in fact an alternative notation for the quantile function of F_Z evaluated at α. In the area of Financial Risk Management, VaR is commonly used to estimate the largest potential loss that might be expected from holding a portfolio over a given period of time at a specified confidence level (Crouhy et al. [4]). For example, if a portfolio has a VaR of $1million at the 95 percent confidence level, then VaR is the cutoff loss such that the probability of losing at least $1million is less than 5 percent over a given time period. So VaR is a measure of risk that summarizes the distribution of returns into a single number. Similarly, in this work, we use this VaR as a tool to examine the anticipated life expectancy of a patient with multiple myeloma from diagnosis until death. As stated in Collet [2], multiple myeloma is a disease characterized by the accumulation of abnormal plasma cell in the bone marrow. Its proliferation within the bone causes pain and the destruction of bone tissue. The condition could be fatal unless treated.

To obtain the distribution of F_Z^{-1} under the setting above, we aggregate simulated returns of X_1 and X_2 associated with the weights, w_1 and w_2. In this work, where X_1 and X_2 respectively represent ST and HB, letting w_1=1 and w_2=0, based on the bivariate Weibull distribution, we get VaR for the survival time, and its procedures are based on a large number of simulated samples generated from copula. For our case, we simulated 5,000,000 samples from each copula to calculate VaR. Since our concern is with longest survival time, VaR is evaluated at the upper tail of the returns distribution of simulated values. Table 2 presents the estimated values of VaR at 90 and 95 percent confidence levels, i.e., 90% (longest) survival time and 95% (longest) survival time, in months, from diagnosis until death from multiple myeloma. Expected shortfall (ES) is also used to examine the anticipated longest survival time. ES is the expectation in excess of VaR, indicating what we expect if an event occurs (Crouhy et al. [4]). ES averages data over all levels greater than or equal to VaR, and so it tells us the average size of the survival time in excess of VaR. In practice, ES is simply obtained by calculating the sample mean of the simulated values above the corresponding VaR. The estimates of ES are displayed in Table 2. For example, in the case of the Clayton copula, which is chosen as the most appropriate copula for data, VaR (95%) and ES (95%) show that the survival times of a patient under treatment could extend to 6.1 and 8.8 years, respectively.

5. CONCLUDING REMARKS

Using Archimedean copulas, bivariate Weibull distributions were constructed. Selecting a copula that may best fit data is important in applications. In an application for multiple myeloma data, it was shown that the Clayton copula best fits the data among the copulas considered. Four different copulas with the different tail dependencies were used to determine this outcome. With extra computing costs, the goodness-of-fit testing procedures of the copula chosen were evaluated. The tail dependence was identified and explained graphically. Based on the bivariate Weibull distribution, we calculated value at risk, where attention is confined to the upper tail of the distribution, to examine the anticipated longest life expectancy of a patient.

6. REFERENCES

[1] D.G. Clayton, “A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence”. Biometrika, 65:141-151, 1978

[2] D. Collect, “Modelling Survival Data in Medical Research”. Chapman, 1999

[3] R.D. Cook and M.E. Johnson, “A family of distributions for modeling non-elliptically symmetric multivariate data”. Journal of the Royal Statistical Society, Series B, 43, 210-218, 1981

[4] M. Crouhy, D. Galai, and R. Mark, “Risk Management”. McGraw-Hill, 2001

[5] B.G. Durie and S.E. Salmon, “A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival”. Cancer, 36(3), 842-854, 1975

[6] V. Durrleman, A. Nikeghbail and T. Roncalli, “Which copula is the right one?”. Credit Lyonnais, Available at SSRN: http://ssm.com/abstract=1032545, 2000

[7] P. Embrechts, A. McNeil and D. Straumann, “Correlation: Pitfall and Alternative”. Risk, 12, 69-71, 1999

[8] M.J. Frank, “On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$”. Aequationes Mathematicae, 19, 194-226, 1979

[9] M.J. Frees and E. Valdez, “Understanding relationships using copulas”. North American Actuarial Journal, 2, 1-25, 1998

[10] C. Genest and R.J. MacKay, “Copules Archimediennes et Failles de Lois Bidimensionnelles Don’t les Marges Sont Donnees”, The Canadian Journal of Statistics, 14, 145-159, 1986a

[11] C. Genest and R.J. MacKay, “The joy of copulas: Bivariate distributions with uniform marginals”. American Statistician, 40, 280-283, 1986b

[12] C. Genest, and L. Rivest, “Statistical inference procedures for bivariate Archimedean copulas”. Journal of American Statistical Association, 88, 1034-1043, 1993

[13] P.R. Greipp, J. San Miguel, B.G. Durie, J.J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J.A. Child, H. Avet-Loiseau, R.A. Kyle, J.J. Lahuerta, H. Ludwig, G. Morgan, R. Powles, K. Shimizu, C. Shustik, P. Sonneveld, P. Tosi, I. Turesson, and J. Westin, “International staging system for multiple myeloma”. Journal of Clinical Oncology, 23, 3412-3420, 2005

[14] E.J. Gumbel, “Bivariate exponential distributions”. Journal of American Statistical Association, 55, 698-707, 1960

[15] P. Hougaard, “A class of multivariate failure time distributions”. Biometrika, 73, 671-678, 1986

[16] H. Joe, “Multivariate Models and Dependence Concepts”. Chapman & Hall, London, 1997

[17] P. Jorion, “Value at Risk: The New Benchmark for Managing Financial Risk”. McGraw-Hill Publication, 2007

[18] J.M. Krall, V.A. Uthoff and J.B. Harley, “A step-up procedure for selecting variables associated with survival”. Biometrics, 31, 49-51, 1975

[19] P. Kumar and M.M. Shoukri, “Evaluating Aortic Stenosis using the Archimedean copula methodology”. Journal of Data Science, 6, 173-187, 2008

[20] R.A. Kyle and S.V. Rajkumar, Multiple myeloma. Blood, 111(6), 2962-2972, 2008

[21] S. Lee, E.-J. Lee and B.O. Omolo, “Using integrated weighted survival difference for the two-sample censored data problem”. Computational Statistics and Data Analysis, 52, 4410-4416, 2008

[22] S. Lee and S. Yang, “Checking the censored two-sample accelerated life model using integrated cumulative hazard difference”. Lifetime Data Analysis, 13, 371-380, 2007

[23] D.Y. Lin, L.J. Wei and Z. Ying, “Checking the cox model with cumulative sums of martingale-based residuals”. Biometrika, 80, 557-72, 1993

[24] A. McNeil, R. Frey and P. Embrechts, “Quantitative Risk Management: Concepts, Techniques and Tools”. Princeton University Press, 2005

[25] M.R. Melchiori, “Which Archimedean copula is the right one?”. Yield Curve, 37, 1-20, 2003

[26] R.B. Nelsen, “An introduction to copulas”. Springer, 1999

[27] L. Rivest and M. Wells, “A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring”. Journal of Multivariate Analysis, 79, 138-155, 2001

[28] A. Sklar, “Functions de repartition a n dimensions et leurs merges”. Publication of the Institute of Statistics, University of Paris, 8, 229-231, 1959

[29] G. Venter, “Tails of copulas”. Proceedings of the Astin Colloquium, 2001.

[30] M. Zheng and J. Klein, “Estimates of marginal survival for dependent competing risks based on an assumed copula”. Biometrika, 82, 127-138, 1995