國立交通大學

統計學研究所

碩士論文

二維存活資料之模式檢驗

Model Diagnostics for Archimedean Copula Models

研 究 生：林建威

指導教授：王維菁 博士

中 華 民 國 九 十 六 年 六 月

二維存活資料之模式檢驗

Model Diagnostics for Archimedean Copula Models

研究生：林建威　　Student: Chien-Wei Lin

指導教授：王維菁 博士　　Advisor: Dr. Wei-Jing Wang

國立交通大學

統計所研究所

碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

二維存活資料之模式檢驗

研究生：林建威　　　　　　　　指導教授：王維菁 博士

國立交通大學統計學研究所

摘要

在本論文中，我們針對右設限的資料提出了 Archimedean Copula(AC)模型的模式檢驗法。我們拓展了 Shih(Biometrika,1998) 的想法，Shih 只針對 Clayton 模型作推論，而我們將之延伸到更大的集合，AC 家族。我們也針對 AC 家族提出了新的資料生成演算法。我們提供模擬分析以佐證在有限樣本下，我們所提出的方法之效能。

關鍵字: Archimedean Copula, Clayton 模型, Concordance 估計量, Frailty 模型, Gumbel 模型

Model Diagnostics for Archimedean Copula Models

Student: Chien-Wei Lin Advisor: Dr. Wei-Jing Wang

Institute of Statistics

National Chiao Tung University

ABSTRACT

In the thesis, we propose a model diagnostic approach to selecting an Archimedean Copula (AC) model based on right censored data. The proposed method extends the idea of Shih (Biometrika, 1998), who considered the Clayton model, to a larger class of models, namely the AC family. We also propose a new algorithm for generating a model from the AC family. Simulation results are provided to examine finite-sample performances of the proposed method.

Keywords: Archimedean Copula, Clayton model, Concordance estimator, Frailty model, Gumbel model.

誌　謝

首先，我要先謝謝王維菁老師在這一年半對我的指導，讓我對於研究該有的態度與精神有所體認，以及教導我在論文的內容陳述上必需的技巧，才使我能夠順利的完成我的碩士論文。雖然在剛開始很不能適應技巧的改變，不斷地修改簡報內容時，實在是受到不小的打擊，但在老師耐心地循循善誘之下，終於有所成果。老師，謝謝您。我還要謝謝江村剛志學長，回想起在剛開始與學長一起做研究時，我非常地笨拙且不得要領，但學長不厭其煩一五一十地教導我基礎觀念，分享他的經驗技巧，幫助我在之後的研究當中，理論推導得以如魚得水。且由於過程以日文對談，我的日文也得以突飛猛進，對學長的感恩實在不是一言兩語說得完。學長在不久的將來就要回日本繼續精進，在這邊我也祝福學長能夠有更進一步的突破，且期許將來有再會的一日。還有所上的教授，謝謝您們，在這兩年的學習歷程中，教授們傳授了非常多難得的觀念以及思維，實在是讓我受益良多，回想起在碩一上陳鄰安老師所開的數統，老師經常掛在嘴邊的一句話，統計，是一門哲學，這句話實在是呼應了這兩年的一切。統計，是門科學，更是門哲學。

再來，我要謝謝一起共度這兩年的好夥伴們：謝謝永在、阿淳、俊睿、柯董、阿Q、益銘、侑侑、小米、小B陪我一起打籃球，讓我再度踏上球場。謝謝小米、俊睿、益銘、益通經常與我分享許多的經驗與想法。謝謝與我一起打CS的好戰友們，陪我度過許多快樂的時光。謝謝平時與我互相討論作業、切磋的好同學們，讓我的思維能夠更加清楚。謝謝燒肉團的成員們，陪我一起大啖美食與聊天。謝謝天天陪我一起共進午晚餐的好同學們，總是配合我的任性到多多吃飯。感謝眾美女們，不嫌棄與我這個大宅男聊天，我會加油的！感謝所有的同學容忍我占了許多電腦室的電腦跑模擬，沒有你們的包容，就沒有我的論文。再一次地謝謝所有的同學，由於個人文筆的拙劣，心中的感觸以及感恩的心情實在難以一一陳述，還請大家見諒了。有你們陪伴的這兩年，我非常地開心，在這裡祝福大家將來都能夠順心如意，不論在事業或學業上，都能夠有更進一步的突破。珍重再見，期望將來有再會的一日。

最後，我要謝謝最親愛的家人，若不是家人的苦心栽培，以及在我煩惱迷惘的時候，

有家人的即時點醒、支持，肯定沒有今日的我。謝謝您們。

在此，將本篇論文獻給我所有的好朋友、師長以及家人，謝謝你們。

林建威　謹誌於

國立交通大學統計研究所

中華民國九十六年六月

Contents

Chapter 1: Introduction

1.1 Motivation

In the literature of survival analysis, there has been substantial research on investigating the association among several lifetime variables. Copula models are the most common modeling choice because they possess nice properties that are suitable for describing lifetime variables.

Specifically copula models form a class of bivariate distributions whose marginals are uniform on the unit interval (Genest and MacKay, 1986). Usually one can write

$$C(u,v) = \Pr(U < u, V < v),$$

where (U,V) are uniform $(0,1)$ variables marginally but correlated with the joint distribution function $C(.,.):[0,1]^2 \to [0,1]$. Let (X,Y) be a pair of continuous failure times. In applications of lifetime data analysis, the copula structure is usually imposed on the survival function such that

$$\Pr(X > x, Y > y) = C\{\Pr(X > x), \Pr(Y > y)\} \qquad (1.1)$$

Models in the copula family allow for separate investigation on the dependence structure and the marginal distributions. The former is often the main interest and hence is handled parametrically (i.e. the form of $C_\alpha(u,v)$ is given). The latter is of less interest and hence dealt with nonparametrically. There has been a trend to derive general properties for a class of models rather than only a single member. The Archimedean copula (AC) family, which is a sub-class of the copula family, is attractive due to its nice analytical properties. For an AC model, the bivariate copula function $C_\alpha(u,v)$ can be further simplified as

$$C_\alpha(u,v) = \phi_\alpha^{-1}\{\phi_\alpha(u) + \phi_\alpha(v)\} \text{ for } u,v \in [0,1], \qquad (1.2)$$

where $\phi_\alpha(\cdot):[0,1] \to [0,\infty]$ is a univariate function which has two continuous derivatives satisfying $\phi_\alpha(1)=0$, $\phi'_\alpha(t)=\dfrac{\partial\phi_\alpha(t)}{\partial t}<0$ and $\phi''_\alpha(t)=\dfrac{\partial^2\phi_\alpha(t)}{\partial t^2}>0$. AC models have the nice feature that the bivariate relationship can be summarized by the univariate function $\phi_\alpha(\cdot)$.

Many authors have considered semi-parametric inference of the copula parameter α, which measures the level of association, without specifying the marginal distributions. Note that α is related to Kendall's τ, a rank correlation measure, such that

$$\tau(\alpha)=4\int_0^1\int_0^1 C_\alpha(u,v)C_\alpha(du,dv)-1, \tag{1.3}$$

where τ is defined as the difference of concordance and discordance probabilities for two independent pairs of (X,Y). These semi-parametric inference procedures require specification of $C_\alpha(.,.)$ or $\phi_\alpha(\cdot)$.

A practical and important question is how can we select an appropriate model to fit the data? There have been some works on model selection including the papers by Genest and Rivest (1993), Shih (1998) and Wang and Wells (2000), just to name a few. In the thesis, we extend the approach of Shih (1998), who considered only testing the Clayton model, to general Archimedean copula models.

1.2 Outline of the thesis

In Chapter 2, we review AC models and their properties. In Chapter 3, we review literature on model diagnostics, including general methodology and results developed for selecting a particular copula model. The proposed method is presented in Chapter 4. In Chapter 5, we review existing data generation algorithms and propose a new approach. Simulation analysis is presented in Chapter 6 and concluding remarks are given in Chapter 7.

Chapter 2: Review of Archimedean Copula models

Under Archimedean Copula family, the relationship between association parameter α and Kendall's τ can be expressed as follows:

$$\tau(\alpha) = 4\int_0^1\int_0^1 C_\alpha(u,v)C_\alpha(du,dv) - 1$$

$$= 4E\left[C_\alpha(u,v)\right] - 1$$

$$= 4\int_0^1 \frac{\phi_\alpha(v)}{\phi'_\alpha(v)}dv + 1. \tag{2.1}$$

If the form of the function $\phi(\cdot)$ is specified, then we can estimate the association parameter α semi-parametrically by the above equation.

In multivariate survival analysis, we usually use Kendall's τ to measure the dependence between random variables. Moreover, we have another dependence measurement called the local odds ratio which is related to the conditional version of Kendall's τ in Oakes (1989). Local odds ratio has been used to measure the pointwise dependence. From Oakes (1989), for an Archimedean Copula model we know that the local odds ratio depends on $t = (x, y)$ only through some function of $S(x, y)$, that is, $\theta^*(x, y) = \theta\{S(x, y)\}$, where $\theta^*(x, y)$ is the local odds ratio function defined as

$$\theta^*(x, y) = \frac{\Pr(X = x, Y = y)\cdot\Pr(X \geq x, Y \geq y)}{\Pr(X = x, Y \geq y)\cdot\Pr(X \geq x, Y = y)} \tag{2.2}$$

For an AC model, we have $\theta(v) = -v\phi''(v)/\phi'(v)$. The paper by Frees and Valdez (1998) provides a nice review of copula models.

We briefly summarize commonly seen members of the AC family.

Example 1: Clayton model

The generating function can be written as $\phi_\alpha(v)=\left(v^{-\alpha}-1\right)\big/\alpha$, $\alpha\in(0,\infty)$. The joint survival function can be written as

$$\Pr(X>x,Y>y)=\left\{\left[S_x(x)\right]^{-\alpha}+\left[S_y(y)\right]^{-\alpha}-1\right\}^{-1/\alpha}.$$

It follows that $\tau=\dfrac{\alpha}{\alpha+2}$.

A special property of the Clayton model reflects in its local odds ratio which can be expressed as $\theta^*(x,y)=\alpha+1$. Notice that $\theta\{S(x,y)\}$ does not depend on (x,y).

Example 2: Gumbel model

The generating function can be written as $\phi_\alpha(v)=\{-\log(v)\}^{\alpha+1}$, $\alpha\in[0,\infty)$. The joint survival function can be written as

$$\Pr(X>x,Y>y)=\exp\left\{-\left[\left(-\ln S_x(x)\right)^{\alpha+1}+\left(-\ln S_y(y)\right)^{\alpha+1}\right]^{\frac{1}{\alpha+1}}\right\}.$$

It follows that $\tau=\dfrac{\alpha}{\alpha+1}$. And the local odds ratio, can be expressed as $\theta^*(x,y)=1-\dfrac{\alpha}{\log S(x,y)}$. Compared with the Clayton model, however $\theta^*(x,y)$ depends on the joint survival function at (x,y).

In Figure 2.1, the curves of $\theta(v)$ for three AC models with the same Kendall's $\tau=0.75$ are plotted. Note that the relationships between α and τ between two models are different. Under the same value of Kendall's τ, the corresponding values of α are different. Note that in the Clayton model, we have $\alpha=\dfrac{2\tau}{1-\tau}=6$. In the Gumbel model, $\alpha=\dfrac{\tau}{1-\tau}=3$. For the Frank model, by setting $\tau=0.75$, we know that $\alpha=7.741\text{e-}07$ which can be obtained by numerically solving

$$\tau=1-4\left\{D_1\left[-\log(\alpha)\right]-1\right\}\big/\log(\alpha)$$

where

$$D_1(\alpha) = \int_0^\alpha \{t/\alpha(e^t - 1)\}\,dt\,.$$

It is worthy to mention that the local odds ratio, $\theta(v)$, plays an important role for our proposed method.

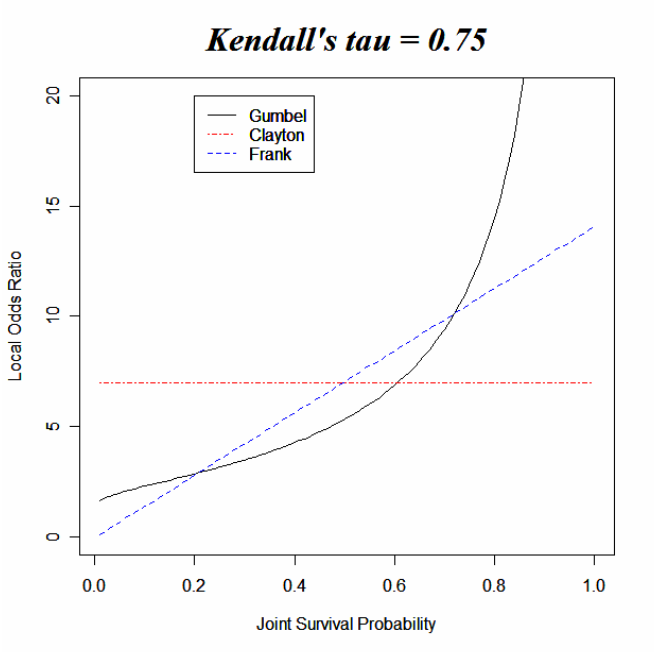


Fig.2.1. The Curves of Local Odds Ratio Function

Chapter 3: Review Methods of Model Checking

In §3.1, we briefly describe methods of model checking for a parametric distribution. In §3.2, we review useful results for selecting an appropriate Archimedean Copula model.

3.1 Model checking for a parametric distribution

Suppose (X,Y) follows a parametric model with the distribution function $F_\theta(x,y)=\Pr(X\le x, Y\le y)$. Usually one can compare $F_{\hat\theta}(x,y)$ with its empirical estimator,

$$\bar F(x,y)=\frac{\sum_{i=1}^{n} I(X_i\le x, Y_i\le y)}{n}.$$

The comparison can be made based on the plots of two curves or some distance measures. For example: one may use the Q-Q plot to check whether the two quantiles are about the same. Alternatively one can set up for a formal hypothesis and test it using statistics such as the K-S test or Chi-squared test, both of which measure the "distance" between the two functions.

3.2 Model selection for Copula models

If the marginal distributions of X and Y were known, the approach mentioned above can be easily applied. Specifically let $U_i=S_1(X_i)$ and $V_i=S_2(Y_i)$, then we have (U_i,V_i) $(i=1,\ldots,n)$, and can compare $C_{\hat\alpha}(u,v)$ and its empirical estimator

$$\bar C(u,v)=\frac{\sum_{i=1}^{n} I(U_i<u, V_i<v)}{n}.$$

However, the parametric forms of the marginal distributions are usually not specified, we have to give up this method.

Genest and Rivest (1993) derived useful properties of AC models which have been applied for model selection. Specifically they define the distribution Copula $V=C\big(F(X),F(Y)\big)$ and find that V is distributed as $K(v)=v-\lambda(v)$, where

$\lambda(v)=\phi(v)/\phi'(v)$ for $0<v\le 1$. Therefore we can compare the difference between a nonparametric estimator of $\Pr\left(F(X,Y)\le v\right)=\Pr(V\le v)=K(v)$ and if model-based representation, $v-\lambda_\alpha(v)$, based on a selected distance measure. If the difference is small, we can say that the imposed model is appropriate for the data.

We briefly illustrate how to perform the above ideas based on complete data $(X_1,Y_1),\ldots,(X_n,Y_n)$. The purpose is to identify the form of ϕ_α. First of all, we need observations $\tilde{V}_i=F(X_i,Y_i)$ in order to estimate $K(v)$ nonparametrically. Since the form of $F(.,.)$ is unknown, Genest and Rivest (1993) proposed the following "pseudo" observations:

$$V_i=\#\left\{(X_j,Y_j): X_j<X_i, Y_j<Y_i\right\}\big/(n-1),\ i=1,\ldots,n,$$

which are proxies of $\tilde{V}_i=F(X_i,Y_i)$. The procedure of model selection is stated below.

1. Obtain the nonparametric estimate of $K(v)$: $K_n(v)=\frac{1}{n}\sum_{i=1}^{n} I(V_i\le v)$.
2. Construct a semi-parametric estimate of $K(v)$. We may have several candidates of models indexed by $\phi_\alpha^{(j)}$ for $j=1,\ldots,J$. For each candidate we need to estimate the value of α. Note that one may estimate Kendall's τ and use the relationship between τ and α to estimate α. Based on

   $$K_\alpha^{(j)}(v)=v-\frac{\phi_\alpha^{(j)}(v)}{\partial\phi_\alpha^{(j)}(v)/\partial v},$$

   we can estimate $K_\alpha^{(j)}(v)$ for $j=1,\ldots,J$.
3. Then we can compare the distance between $K_n(v)$ and $K_\alpha^{(j)}(v)$ for $j=1,\ldots,J$. The most-fitted model is the one which gives the smallest distance between the two curves.

The above procedure is not applicable when there is censoring. Wang and Wells (2000) propose a nonparametric estimator of $K(v)$ based on right censored data.

Chapter 4: The proposed method for model checking

In this chapter, we present our proposal for model checking. The idea was motivated by the paper of Shih (1998) who proposed to test the Clayton model by comparing the difference between weighted and unweighted concordance estimators of the association parameter α. When the model assumption is correct, which is the condition of the null hypothesis, both estimators converge to the true parameter value. On the other hand, when the model assumption is false, the two estimators will converge to different values. Here we extend Shih's idea to verify whether the model follows a particular Archimedean Copula model. We will use the Gumbel model as an example of AC models.

Define $\Delta_{ij} = I\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right]$, where $\left(X_i, Y_i\right)$ and $\left(X_j, Y_j\right)$ are two independent replications of $\left(X, Y\right)$. This indicator variable denotes whether the pairs $\left(X_i, Y_i\right)$ and $\left(X_j, Y_j\right)$ are concordant or discordant. The conditional expectation of Δ_{ij} contains the information about the level of association. Specifically it follows that

$$\Pr\left(\Delta_{ij}=1\middle|\tilde{X}_{ij} = x, \tilde{Y}_{ij} = y\right)$$

$$= \frac{2 \cdot \Pr\left(X_i = x, Y_i = y\right) \cdot \Pr\left(X_j \geq x, Y_j \geq y\right)}{2 \cdot \left[\Pr\left(X_i = x, Y_i = y\right) \cdot \Pr\left(X_j \geq x, Y_j \geq y\right) + \Pr\left(X_i = x, Y_i \geq y\right) \cdot \Pr\left(X_j \geq x, Y_j = y\right)\right]}$$

$$= \frac{\dfrac{\Pr\left(X = x, Y = y\right) \cdot \Pr\left(X \geq x, Y \geq y\right)}{\Pr\left(X = x, Y \geq y\right) \cdot \Pr\left(X \geq x, Y = y\right)}}{\dfrac{\Pr\left(X = x, Y = y\right) \cdot \Pr\left(X \geq x, Y \geq y\right)}{\Pr\left(X = x, Y \geq y\right) \cdot \Pr\left(X \geq x, Y = y\right)} + 1}$$

$$= \frac{\theta_\alpha\left(S\left(x, y\right)\right)}{\theta_\alpha\left(S\left(x, y\right)\right) + 1}, \qquad (4.1)$$

where $\tilde{X}_{ij} = \min\left(X_i, X_j\right)$ and $\tilde{Y}_{ij} = \min\left(Y_i, Y_j\right)$ and $\theta_\alpha\left(S\left(x, y\right)\right)$ is the local odds ratio defined in equation (2.2). Recall that for the Clayton model, $\theta_\alpha\left(S\left(x, y\right)\right) = \alpha + 1$ and the Gumbel model,

$$\theta_\alpha\left(S(x,y)\right)=1-\frac{\alpha}{\log S(x,y)}.$$

Now we will illustrate how to utilize equation (4.1) to construct different forms of estimating functions of α. In § 4.1, external censoring is ignored temporarily to simply the presentation. In § 4.2, the proposed methods are modified to handle censored data.

4.1: Analysis based on Complete Data

For complete data, we observe $\left\{(X_i,Y_i)(i=1,\ldots,n)\right\}$ which is a random sample of (X,Y). Based on the moment condition of equation (4.1), one can construct the following estimating function of the association parameter α:

$$U_0\left(\alpha,S(x,y)\right)=\sum_{i<j}\left[\Delta_{ij}-E\left(\Delta_{ij}\mid\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right],$$

where $E\left(\Delta_{ij}\mid\tilde{X}_{ij},\tilde{Y}_{ij}\right)=\Pr\left(\Delta_{ij}=1\middle|\tilde{X}_{ij},\tilde{Y}_{ij}\right)=\dfrac{\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)+1}.$

For the Gumbel model, the above estimating function becomes

$$U_0\left(\alpha,S(x,y)\right)=\sum_{i<j}\left[\Delta_{ij}-\frac{1-\dfrac{\alpha}{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}}{2-\dfrac{\alpha}{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}}\right]$$

$$=\sum_{i<j}\left[\Delta_{ij}-\frac{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha}{2\cdot\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha}\right]. \qquad (4.2)$$

Here notice that, we use the conditional probability rather than the unconditional one. The latter is used in Shih's paper (1998) since both are equivalent under the Clayton model.

Note that $S(x,y)=\Pr(X>x,Y>y)$ is a nuisance function. For complete data, $S(x,y)$ can be estimated by the empirical estimator

$$\hat{S}(x,y)=\frac{\sum_{i=1}^{n}I\left(X_i>x,Y_i>y\right)}{n}.$$

To obtain an estimator of α, $\hat{\alpha}$, we solve $U_0\left(\alpha, \hat{S}(x,y)\right)=0$. Since an explicit solution is not available, we suggest to solve the equation by numerical methods, say the Newton-Raphson method, which often requires computing the derivative of $U_0\left(\alpha, S(x,y)\right)$ with respect to α. For the Gumbel model, the derivative equals

$$\sum_{i<j}\left[\frac{-\dfrac{d\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}}{\left(1+\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right)^2}\right]=\sum_{i<j}\left[\frac{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}{\left[2\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]^2}\right].$$

The function in (4.2) can be viewed as an unweighted version. That is, Δ_{ij} is treated equally for each combination of i and j. However since different combinations of Δ_{ij} are associated with different values of (X_i, Y_i) and (X_j, Y_j), it is reasonable to suspect that such additional information may be utilized in the estimation procedure. Clayton (1978) proposed a conditional likelihood function for the Clayton family. We can modify his method for AC models. The resulting log-likelihood function is given by

$$\mathrm{L}(\alpha)=\sum_i \log\left(\frac{\theta_\alpha\left(S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)\right)}{R_{ii}-1+\theta_\alpha\left(S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)\right)}\right)+\sum_{i<j}\left(1-\Delta_{ij}\right)\log\left(\frac{R_{ij}-1}{R_{ij}-1+\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}\right),$$

where $R_{ij}=R\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)$ and $R(x,y)=\sum_{i=1}^{n} I\left(X_i \ge x, Y_i \ge y\right)$. Now we derive the score function for an Archimedean copula model:

$$\frac{\partial L(\alpha)}{\partial \alpha}=\sum_i \frac{\left(R_{ii}-1\right)}{\left[R_{ii}-1+\theta_\alpha\left(S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)\right)\right]}\cdot\frac{d\ln\theta_\alpha\left(S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)\right)}{d\alpha}+\sum_{i<j}\frac{\left(\Delta_{ij}-1\right)}{\left[R_{ij}-1+\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right]}\cdot\frac{d\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}.$$

For the Gumbel model, the score function becomes

$$\frac{\partial L(\alpha)}{\partial \alpha}=\sum_i \frac{-\left(R_{ii}-1\right)}{\left(R_{ii}-\dfrac{\alpha}{\log S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)}\right)\left[\log S\left(\tilde{X}_{ii},\tilde{Y}_{ii}\right)-\alpha\right]}+\sum_{i<j}\frac{\left(1-\Delta_{ij}\right)}{\left[R_{ij}\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}.$$

The two terms in the right-hand side of the above equation can be combined using the technique in Oakes (1986). That is, if $j \in R_{ii}$, then $R_{ii} = R_{ij}$, $\Delta_{ij} = 1$ and $\sum_{j \in R_{ii}} \Delta_{ij} = R_{ii} - 1$.

It follows that

$$
\begin{aligned}
S(\alpha) &= \frac{\partial L(\alpha)}{\partial \alpha} \\
&= \sum_{i} \frac{\sum_{j \in R_{ii}} \Delta_{ij}}{\left[R_{ii} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ii}, \tilde{Y}_{ii}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ii}, \tilde{Y}_{ii}\right)\right)}{d\alpha} + \sum_{i<j} \frac{\left(\Delta_{ij} - 1\right)}{\left[R_{ij} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} \\
&= \sum_{i} \sum_{j \in R_{ii}} \frac{\Delta_{ij}}{\left[R_{ii} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ii}, \tilde{Y}_{ii}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ii}, \tilde{Y}_{ii}\right)\right)}{d\alpha} + \sum_{i<j} \frac{\left(\Delta_{ij} - 1\right)}{\left[R_{ij} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} \\
&= \sum_{i<j} \frac{\Delta_{ij}}{\left[R_{ij} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} + \frac{\left(\Delta_{ij} - 1\right)}{\left[R_{ij} - 1 + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} \\
&= \sum_{i<j} \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right) + 1}{\left[\left(R_{ij} - 1\right) + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} \left\{\Delta_{ij} - \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right) + 1}\right\}.
\end{aligned}
\tag{4.3}
$$

For the Gumbel model for illustration, the equation equals

$$
S(\alpha) = -\sum_{i<j} \frac{2\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha}{\left[R_{ij} \log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha\right]} \cdot \frac{\left[\Delta_{ij} - \dfrac{\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha}{2\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha}\right]}{\left[\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha\right]}.
$$

If we treat the estimating function in (4.2) as an unweighted version, the weight function in equation (4.3) is

$$
\frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right) + 1}{\left[\left(R_{ij} - 1\right) + \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha},
$$

which accounts for the effects of the data and model properties.

Next, one can derive the Fisher information function, that is, the derivative of minus score function. It follows that

$$I(\alpha) = -\frac{\partial S(\alpha)}{\partial \alpha}$$

$$= -\frac{\partial}{\partial \alpha} \sum_{i<j} \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)+1}{\left[\left(R_{ij}-1\right)+\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha} \left\{ \Delta_{ij} - \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)+1} \right\}$$

$$= -\sum_{i<j} \frac{\left[\dfrac{d\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha}\right]^2}{\left[R_{ij}-1+\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} \left\{ \left[1 - \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)+1}{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)} \Delta_{ij} \right] \cdot \right.$$

$$\left. \left[\frac{1}{\left[R_{ij}-1+\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)\right]} - \frac{\dfrac{d^2\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha^2}}{\left[\dfrac{d\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{d\alpha}\right]^2} \right] - \frac{\Delta_{ij}}{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)^2} \right\}.$$

For the Gumbel model, the Fisher information function becomes,

$$I(\alpha) = -\sum_{i<j} \frac{\left\{ \dfrac{\left[2\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)-\alpha\right]\left(1-\Delta_{ij}\right)-\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\left[R_{ij}\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)-\alpha\right]} - \dfrac{\Delta_{ij}\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\left[\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)-\alpha\right]} \right\}}{\left[R_{ij}\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)-\alpha\right]\left[\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)-\alpha\right]}.$$

Again the nuisance function $S(x, y)$ can be estimated separately and then the estimator is plugged into the score function and Fisher information function. The solution $\hat{\alpha}_w$ requires using numerical methods such as the Newton-Raphson method approach.

4.2: Analysis based on Right Censored Data

Now we incorporate external situation censoring in the analysis. Let (A_i, B_i) be the bivariate censoring variables. One only observes $\left(\tilde{X}_i, \tilde{Y}_i, \delta_{1i}, \delta_{2i}\right)$, where $\tilde{X}_i = \min(X_i, A_i)$, $\tilde{Y}_i = \min(Y_i, B_i)$, $\delta_{1i} = I(X_i \le A_i)$ and $\delta_{2i} = I(Y_i \le B_i)$. In presence of right censoring, we know the order of X_i and X_j if and only if $\tilde{X}_{ij} \le \min(A_i, A_j)$. Similarly the order of Y_i

and Y_j can be known if and only if $\tilde{Y}_{ij} \le \min(B_i, B_j)$. Define $Z_{ij} = I\left(\tilde{X}_{ij} \le \tilde{A}_{ij}, \tilde{Y}_{ij} \le \tilde{B}_{ij}\right)$, where $\tilde{A}_{ij} = \min(A_i, A_j)$ and $\tilde{B}_{ij} = \min(B_i, B_j)$. The value of Δ_{ij} is observed if and only if $Z_{ij} = 1$. We will modify the estimating procedures that only include comparable pairs (i.e. those with $Z_{ij} = 1$) in the analysis.

The unweighted estimating function of the association parameter α can be modified as

$$U_0\left(\alpha, S(x, y)\right) = \sum_{i<j}\left[\Delta_{ij} - E\left(\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right] \cdot Z_{ij},$$

where

$$E\left(\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}\right) = \Pr\left(\Delta_{ij} = 1 \middle| \tilde{X}_{ij}, \tilde{Y}_{ij}\right) = \frac{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right)}{\theta_\alpha\left(S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right) + 1}.$$

For the Gumbel model, the above function then becomes

$$U_0\left(\alpha, S(x, y)\right) = \sum_{i<j}\left[\Delta_{ij} - \frac{1 - \dfrac{\alpha}{\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}}{2 - \dfrac{\alpha}{\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}}\right] \cdot Z_{ij}$$

$$= \sum_{i<j}\left[\Delta_{ij} - \frac{\log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha}{2 \cdot \log S\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) - \alpha}\right] \cdot Z_{ij}.$$

For estimating the function $S(x, y)$ for right censored data, there exist several nonparametric estimators. Here we adopt the Dabrowska estimator (1988) which is the most well-known one among its competitors. The formula of $\hat{S}(x, y)$ is given by

$$\hat{S}(x, y) = \hat{S}(x, 0)\hat{S}(0, y) \prod_{\substack{0<u\le x \\ 0<v\le y}} \left[1 - L(\Delta u, \Delta v)\right],$$

where

$$\hat{L}(\Delta u, \Delta v) = \frac{\hat{\Lambda}_{10}(\Delta u, v-)\hat{\Lambda}_{01}(u-, \Delta v) - \hat{\Lambda}_{11}(\Delta u, \Delta v)}{\left\{1 - \hat{\Lambda}_{10}(\Delta u, v-)\right\}\left\{1 - \hat{\Lambda}_{01}(u-, \Delta v)\right\}},$$

$$\hat{\Lambda}_{11}\left(\Delta u,\Delta v\right)=\frac{\sum_{i=1}^{n}I\left(X_i=u,Y_i=v,\delta_{1i}=1,\delta_{2i}=1\right)}{\sum_{i=1}^{n}I\left(X_i\geq u,Y_i\geq v\right)},$$

$$\hat{\Lambda}_{10}\left(\Delta u,v-\right)=\frac{\sum_{i=1}^{n}I\left(X_i=u,Y_i\geq v,\delta_{1i}=1\right)}{\sum_{i=1}^{n}I\left(X_i\geq u,Y_i\geq v\right)},$$

$$\hat{\Lambda}_{01}\left(u-,\Delta v\right)=\frac{\sum_{i=1}^{n}I\left(X_i\geq u,Y_i=v,\delta_{2i}=1\right)}{\sum_{i=1}^{n}I\left(X_i\geq u,Y_i\geq v\right)},$$

and $\hat{S}\left(x,0\right)$ and $\hat{S}\left(0,y\right)$ are the usual Kaplan-Meier estimates, i.e.,

$$\hat{S}\left(x,0\right)=\prod_{u\leq x}\left\{1-\frac{\sum_{i=1}^{n}I\left(X_i=u,\delta=1\right)}{\sum_{i=1}^{n}I\left(X_i\geq u\right)}\right\},$$

$$\hat{S}\left(0,y\right)=\prod_{u\leq y}\left\{1-\frac{\sum_{i=1}^{n}I\left(Y_i=u,\delta=1\right)}{\sum_{i=1}^{n}I\left(Y_i\geq u\right)}\right\}$$

That is, the marginals of $\hat{S}\left(x,y\right)$ are given by the univariate Kaplan-Meier estimates.

Numerical algorithms for solving $\hat{\alpha}$ often involve calculating the derivative of $U_0\left(\alpha,S\left(x,y\right)\right)$. For Gumbel's model, this terms equals

$$\sum_{i<j}\left[\frac{-\dfrac{d\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}}{\left(1+\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right)^2}\right]\cdot Z_{ij}=\sum_{i<j}\left[\frac{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}{\left[2\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]^2}\right]\cdot Z_{ij}.$$

The weighted versions of the score function and Fisher information for censored data are given by

$$S\left(\alpha\right)=\sum_{i<j}\frac{\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)+1}{\left[\left(R_{ij}-1\right)+\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right]}\cdot\frac{d\ln\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}\left\{\Delta_{ij}-\frac{\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{\theta_\alpha\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)+1}\right\}\cdot Z_{ij}.$$

For the Gumbel model, it equals

$$S(\alpha)=-\sum_{i<j}\frac{2\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha}{\left[R_{ij}\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}\cdot\frac{\left[\Delta_{ij}-\dfrac{\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha}{2\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha}\right]}{\left[\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}\cdot Z_{ij}.$$

And

$$I(\alpha)=-\sum_{i<j}\frac{\left[\dfrac{d\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}\right]^{2}}{\left[R_{ij}-1+\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right]}\left\{\left[1-\frac{\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)+1}{\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}\Delta_{ij}\right]\cdot\right.$$

$$\left.\left[\frac{1}{\left[R_{ij}-1+\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)\right]}-\frac{\dfrac{d^{2}\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha^{2}}}{\left[\dfrac{d\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)}{d\alpha}\right]^{2}}\right]-\frac{\Delta_{ij}}{\theta_{\alpha}\left(S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)\right)^{2}}\right\}\cdot Z_{ij}.$$

For the Gumbel model, it equals

$$I(\alpha)=-\sum_{i<j}\frac{\left\{\dfrac{\left[2\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]\left(1-\Delta_{ij}\right)-\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}{\left[R_{ij}\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}-\dfrac{\Delta_{ij}\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)}{\left[\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}\right\}}{\left[R_{ij}\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]\left[\log S\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)-\alpha\right]}\cdot Z_{ij}.$$

The solution $\hat{\alpha}_w$ requires using numerical methods such as the Newton-Raphson method approach.

4.3: Model checking

By defining $\hat{\gamma}=\log\hat{\alpha}$ and $\hat{\gamma}_w=\log\hat{\alpha}_w$, Shih (1998) proves that when the Clayton model is correct, $n^{\frac{1}{2}}\left(\hat{\gamma}_w-\hat{\gamma}\right)$ converges to a normal distribution with mean zero. For complete data, the variance is

$$W(\eta)=V(\eta)+V_w(\eta)-2H(\eta),$$

where $\eta=1/\alpha$ and

$$V(\eta)=\left\{\frac{8(2\eta+1)^4}{(\eta+1)^2}L(\eta)-\frac{4(2\eta+1)^2(17\eta^3+27\eta^2+14\eta+2)}{3\eta^2(\eta+1)^2(3\eta+1)}\right\}$$

is the variance of $n^{\frac{-1}{2}}\hat{\gamma}$, where $L(\eta)=\sum_{k=0}^{\infty}\frac{\Gamma(3\eta)\prod_{i=1}^{k}(\eta+i)}{\Gamma(k+3\eta+1)(2\eta+k)}$, and

$$V_w(\eta)=2\eta^2+6\eta+5-(\eta+1)^4\left\{\psi'\left(\frac{1}{2}+\frac{1}{2}\eta\right)-\psi'\left(1+\frac{1}{2}\eta\right)\right\}$$

is the variance of $n^{\frac{-1}{2}}\hat{\gamma}_w$, where $\psi'(c)=\sum(n+c)^{-2}$ is the trigamma function, and

$$H(\eta)=-4\eta-\frac{40\eta^3+49\eta^2+21\eta+3}{\eta^2(2\eta+1)}+4J(\eta)(\eta+1)(2\eta+1)^2,$$

where

$$J(\eta)=\sum_{k=0}^{\infty}\frac{k!\Gamma(2\eta)}{(\eta+k)(2\eta+k)\Gamma(2\eta+k)}.$$

For right censored data, the asymptotic variance becomes

$$\tilde{W}(\eta)=\tilde{V}(\eta)+\tilde{V}_w(\eta)-2\tilde{H}(\eta),$$

where

$$\tilde{V}(\eta)=\frac{4(\alpha+2)^2}{(\alpha+1)^2\delta^{*2}}\iint N_{12}N_{13}C(u_2,v_2)C(u_3,v_3)C(u_{23},v_{23})\prod_{i=1}^{3}f(x_i,y_i)dxdy,$$

$$\tilde{V}_w(\eta)=\frac{1}{\delta^2}\iint\frac{N_{12}N_{13}C(u_2,v_2)C(u_3,v_3)C(u_{23},v_{23})}{S(u_2,v_2)S(u_3,v_3)}\prod_{i=1}^{3}f(x_i,y_i)dxdy,$$

$$\tilde{H}(\eta)=\frac{2(\alpha+2)}{(\alpha+1)\delta\delta^*}\iint\frac{N_{12}N_{13}C(u_2,v_2)C(u_3,v_3)C(u_{23},v_{23})}{S(u_2,v_2)}\prod_{i=1}^{3}f(x_i,y_i)dxdy,$$

with $\delta=\Pr(\text{an observation is uncensored in both components})$, $N_{1l}=\Delta_{1l}(\alpha+2)-(\alpha+1)$, $\delta^*=\Pr(Z_{ij}=1)$, $u_l=\min(x_1,x_l)$, $v_l=\min(y_1,y_l)$, $l=2,3$, and $u_{23}=\max(u_2,u_3)$, $v_{23}=\max(v_2,v_3)$.

One can estimating the variance for complete data $W(\eta)$ by $W(\hat{\eta}_w)$. The censored version $\tilde{W}(\eta)$ can be estimated by

$$\hat{W} = \frac{1}{n}\sum_{i\neq j\neq k} Z_{ij}Z_{ik}\hat{N}_{ij}\hat{N}_{ik}\left\{\frac{1}{\tilde{R}_{ij}\hat{\delta}} - \frac{2(\hat{\alpha}_w+2)}{n(\hat{\alpha}_w+1)\hat{\delta}^*}\right\}\left\{\frac{1}{\tilde{R}_{ik}\hat{\delta}} - \frac{2(\hat{\alpha}_w+2)}{n(\hat{\alpha}_w+1)\hat{\delta}^*}\right\},$$

where

$$\hat{N}_{ij} = \Delta_{ij}(\hat{\alpha}_w+2)-(\hat{\alpha}_w+1),\quad \hat{\delta} = \frac{\sum_{i=1}^{n} I(\delta_{1i}=1,\delta_{2i}=1)}{n},\quad \hat{\delta}^* = \sum_{i<j} Z_{ij}\Big/\binom{n}{2}.$$

For complete data, the null hypothesis is rejected when $\dfrac{|\hat{\gamma}_w-\hat{\gamma}|}{\left[W(\hat{\eta}_w)/n\right]^{1/2}}$ is greater than $Z_{1-\alpha/2}$ with significance level equals to α, where Z_p is the p-th percentile of the standard normal distribution. For censored data, the test statistic is changed to $\dfrac{|\hat{\gamma}_w-\hat{\gamma}|}{\left[\hat{W}/n\right]^{1/2}}$.

For our proposal which can be extended to the whole AC family, we need to know the (asymptotic) distribution of $\hat{\gamma}_w-\hat{\gamma}$. Asymptotic normality should be correct based on the central limit theorem and the delta method. Formal derivations will be future work. However the proposed method involves the complicated plugged-in estimator, analytic estimation of the variance term will be impossible. Note that even there exists no analytic form for the variance of Dabrowska's estimator. Hence we suggest to use the Jackknife algorithm to estimate the variance of the proposed test statistic, denoted by $\hat{\sigma}^2_{Jackknife}$. Our hypothesis testing is rejected if $\dfrac{|\hat{\gamma}_w-\hat{\gamma}|}{\hat{\sigma}_{Jackknife}}$ is greater than $Z_{1-\alpha/2}$ with the significance level equals to α.

Chapter 5: Data Generation Algorithms

In this chapter, we discuss two existing algorithms for generating an AC model and then propose a new data generation algorithm.

5.1 Frailty Approach

5.1.1 Theoretical Background

Suppose that there are p lifetime variables, $X_1, X_2, \ldots, X_p$ which are correlated. Oakes (1989) might be the first one who used the idea of frailty to construct multivariate distributions. He assumes that the dependence among these variables can be fully explained by a latent variable γ, called "frailty". That is, given the value of γ, these variables are independent such that one can write

$$S\left(X_1 > x_1, \ldots, X_p > x_p \mid \gamma\right) = \prod_{j=1}^{p} S\left(X_j > x_j \mid \gamma\right).$$

If the failure times represent the lifetimes of family members, γ represents the shared genetic/environmental factor. Furthermore, γ affects each of T_j via a proportional hazard model such that

$$h_j\left(x \mid Z\right) = h_j\left(x\right) \cdot \gamma \text{ (or equivalently } S_j\left(x \mid Z\right) = B_j\left(x\right)^{\gamma}\text{),}$$

where $B_j\left(x\right) = \exp\left(-\int_o^x b_j\left(t\right) dt\right)$. Since γ is a (positive) random variable, the unconditional joint survival function can be expressed as

$$S\left(X_1 > x_1, \ldots, X_p > x_p\right) = E_\gamma\left\{\left[\prod_{j=1}^{p} B_j\left(x_j\right)\right]^{\gamma}\right\}. \qquad (5.1)$$

Notice that the Laplace transform of γ is defined as

$$L\left(t\right) = E_\gamma\left(e^{-t\gamma}\right) = \int e^{-tx} dF_\gamma\left(x\right),$$

where F_γ is the distribution function of γ.

Oakes (1989) also pointed out that there is a relationship between the above frailty family and the AC family introduced earlier. That is the inverse of the generating function $\phi_\alpha^{-1}(\cdot)$ for an AC model is actually the Laplace transform of γ. To see this, we can view $L(t)$ as the moment generating function evaluated at $-t$. If we know the form of $L(t)$, we derive its distribution. Moreover,

$$\begin{aligned} S\left(X_1 > x_1, \ldots, X_p > x_p\right) &= E_\gamma\left\{\left[B_1(x_1)\cdots B_p(x_p)\right]^\gamma\right\} \\ &= E_\gamma\left\{\exp\left[\gamma\cdot\left(\ln B_1(x_1) + \ln B_2(x_2) + \cdots + B_p(x_p)\right)\right]\right\} \\ &= L\left[-\ln B_1(x_1) - \ln B_2(x_2) - \cdots - \ln B_p(x_p)\right]. \end{aligned}$$

Since

$$S_i(x_i) = E_\gamma\left[B_i(x_i)^\gamma\right] = L\left[-\ln B_i(x_i)\right] \Rightarrow -\ln B_i(x_i) = L^{-1}\left[S_i(x_i)\right],$$

we can obtain

$$S\left(X_1 > x_1, \ldots, X_p > x_p\right) = L\left\{L^{-1}\left[S_1(x_1)\right] + \cdots + L^{-1}\left[S_p(x_p)\right]\right\}.$$

Since the Laplace forms have well-defined inverses, thus from the above equation we can find that the inverse function of L, L^{-1}, acts as the generator of Archimedean Copula. That is, the frailty family can be treated as a subclass of the Archimedean copula family with the generator being the inverse function of the Laplace transform for the latent variable γ. The explanation by using the frailty variable to explain the cause of dependence is intuitive for many applications.

5.1.2 Generation Algorithm

Here we consider the bivariate case with $p = 2$ and, to unify the notations, we let $X = X_1$ and $Y = X_2$. Based on the construction of the frailty model, a random replication of $(X, Y) = (X_1, X_2)$ can be generated as follows.

1. Generate a positive random variable γ following a given distribution. Then derive the form of its Laplace transform denoted as L.

2. Independent of γ, generate (U_1, U_2) which are independent uniform $(0,1)$ random variables. Recall that based on the relationship $U_k = S_k(X_k \mid \gamma) = B_k(X_k)^{\gamma}$, we have $-\gamma^{-1} \cdot \ln U_k = -\ln B_k(X_k)$ for $k = 1, 2$.

3. After specifying the forms of $S_k(.)$ $(k = 1, 2)$, we need to find

$$X_k = S_k^{-1}\left[L\left(-\gamma^{-1} \cdot \ln U_k\right)\right].$$

For the Gumbel model, it corresponds to the case that γ following positive stable distribution with Laplace transform

$$L(t) = \exp\left(-t^{\frac{1}{\alpha+1}}\right),$$

where $L^{-1}(t) = \left[-\log(t)\right]^{\alpha+1} = \phi(t)$.

5.2 Conditional Distribution Approach

5.2.1 Theoretical Background

The idea was proposed by Lee (1993). Given the marginal distribution of X_1 and if the conditional distribution of $X_2 \mid X_1$ is specified, then X_2 can be generated. In general, X_k can be generated given that the form of $X_k \mid X_1, X_2, ..., X_{k-1}$ is specified. The algorithm can be performed successively for $k = 2, ..., p$.

Now we apply the above idea to the family of Archimedean copula construction of the form:

$$F(x_1, x_2, ..., x_p) = C\left(F_1(x_1), F_2(x_2), ..., F_p(x_p)\right) = \phi^{-1}\left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\}.$$

The joint distribution function is given by

$$\begin{aligned} f(x_1, \ldots, x_k) &= \frac{\partial^k}{\partial x_1 \ldots \partial x_k} F(x_1, ..., x_k) \\ &= \frac{\partial^k}{\partial x_1 \ldots \partial x_k} \phi^{-1}\left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\} \end{aligned}$$

$$=\left(\phi^{-1}\right)^{(k)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x_k\right)\right]\right\}\cdot\prod_{i=1}^{k}\phi'\left[F_i\left(x_i\right)\right]F_i'\left(x_i\right),$$

where the superscript notation (k) means the k-th derivative. Then, the conditional density function of X_k given $X_1,X_2,...,X_{k-1}$ is

$$f\left(x_k \mid x_1,...,x_{k-1}\right)=\frac{f\left(x_1,\ldots,x_k\right)}{f\left(x_1,\ldots,x_{k-1}\right)}$$

$$=\frac{\left(\phi^{-1}\right)^{(k)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x_k\right)\right]\right\}\cdot\prod_{i=1}^{k}\phi'\left[F_i\left(x_i\right)\right]F_i'\left(x_i\right)}{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]\right\}\cdot\prod_{i=1}^{k-1}\phi'\left[F_i\left(x_i\right)\right]F_i'\left(x_i\right)}$$

$$=\frac{\left(\phi^{-1}\right)^{(k)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x_k\right)\right]\right\}}{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]\right\}}\cdot\phi'\left[F_k\left(x_k\right)\right]F_k'\left(x_k\right).$$

Then, the conditional cumulative density function of X_k given $X_1,X_2,...,X_{k-1}$ is

$$F\left(x_k \mid x_1,...,x_{k-1}\right)=\int_{-\infty}^{x_k} f\left(x \mid x_1,...,x_{k-1}\right)dx$$

$$=\int_{-\infty}^{x_k}\frac{\left(\phi^{-1}\right)^{(k)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x\right)\right]\right\}}{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]\right\}}\cdot\phi'\left[F_k\left(x\right)\right]F_k'\left(x\right)dx$$

$$=\left.\frac{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x\right)\right]\right\}}{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]\right\}}\right|_{x=-\infty}^{x_k}$$

$$=\frac{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_k\left(x_k\right)\right]\right\}}{\left(\phi^{-1}\right)^{(k-1)}\left\{\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]\right\}}$$

$$=\frac{\left(\phi^{-1}\right)^{(k-1)}\left\{a_{k-1}+\phi\left[F_k\left(x_k\right)\right]\right\}}{\left(\phi^{-1}\right)^{(k-1)}\left(a_{k-1}\right)}, \tag{5.2}$$

where $a_{k-1}=\phi\left[F_1\left(x_1\right)\right]+\cdots+\phi\left[F_{k-1}\left(x_{k-1}\right)\right]$.

5.2.2 Generation Algorithm

Consider the bivariate case with $\left(X,Y\right)=\left(X_1,X_2\right)$. The algorithm can be described as following:

1. Generate $\left(U_1,U_2\right)$ independent uniform$\left(0,1\right)$ random variables.

2. Let $X_1 = F_1^{-1}(U_1)$.

3. Then set X_2 as the solution of the following equation:

$$U_2 = \frac{(\phi^{-1})'\{\phi[F_1(x_1)]+\phi[F_2(x_2)]\}}{(\phi^{-1})'\{\phi[F_1(x_1)]\}} = \frac{(\phi^{-1})'\{\phi(U_1)+\phi[F_2(x_2)]\}}{(\phi^{-1})'[\phi(U_1)]}.$$

For the Gumbel model with $\phi(v) = \{-\log(v)\}^{\alpha+1}$, we obtain

$$\phi'(v) = (\alpha+1)[-\log(v)]^{\alpha} \cdot \frac{-1}{v}, \quad \phi^{-1}(v) = \exp(-v^{1/\alpha+1}),$$

$$[\phi^{-1}(v)]' = -\frac{v^{\frac{-\alpha}{\alpha+1}}}{\alpha+1}\exp\left(-v^{\frac{1}{\alpha+1}}\right),$$

then

$$U_2 = \frac{\left\{[-\log(U_1)]^{\alpha+1}+[-\log(F_2(x_2))]^{\alpha+1}\right\}^{-\frac{\alpha}{\alpha+1}} \cdot \exp\left\{-\left\{[-\log(U_1)]^{\alpha+1}+[-\log(F_2(x_2))]^{\alpha+1}\right\}^{\frac{1}{\alpha+1}}\right\}}{U_1[-\log(U_1)]^{-\alpha}}.$$

Obviously, the above form does not allow an explicit solution. Hence to solve the equation, we need to do it numerically.

5.3: The Proposed Data Generation Method

The idea is based on a theorem in Genest & Rivest (1993). Briefly speaking, for (X,Y) which follow an AC model, we can define two random variables (U,V) where

$$U = \phi(S_x(X))/\{\phi(S_x(X))+\phi(S_y(Y))\}$$

and

$$V = S(X,Y) = \phi^{-1}\{\phi(S_x(X))+\phi(S_y(Y))\}.$$

It follows that U is distributed as uniform $(0,1)$,

$$K(v) = \Pr(S(X,Y) \le v) = v - \phi(v)/\phi'(v),$$

and $U \perp V$. These theoretical results can be applied to generate a random replication of

(X,Y) which follows an AC model. The algorithm can be stated as follows.

1. Generate two independent random variables U and U^*, both of which follow a uniform$(0,1)$ distribution.

2. Given an AC model, we can derive the formula of $K(v)$. Then we can obtain $V = S(X,Y)$ by solving $V = K^{-1}(U^*)$. Note that $K(v)$ is a distribution function and hence is monotone increasing. It is easy to find the inverse function $K^{-1}(\cdot)$ numerically to obtain V.

3. Based on the theorem, U and V are independent, where

$$U = \frac{\phi(S_x(X))}{\phi(S_x(X)) + \phi(S_y(Y))}.$$

Since $V = \phi^{-1}\{\phi(S_x(X)) + \phi(S_y(Y))\}$, we have $S_x(X) = \phi^{-1}[U \cdot \phi(V)]$ and $S_y(Y) = \phi^{-1}[(1-U)\cdot\phi(V)]$. Finally we can set

$$X = F_x^{-1}\{1 - \phi^{-1}[U \cdot \phi(V)]\}$$

and

$$Y = F_y^{-1}\{1 - \phi^{-1}[(1-U)\cdot\phi(V)]\},$$

where the forms of $S_x(\cdot)$ and $S_y(\cdot)$ should be specified beforehand.

For the Gumbel model with $\phi(v) = \{-\log(v)\}^{\alpha+1}$, we have

$$\phi^{-1}(v) = \exp\left(-v^{1/\alpha+1}\right),\ \phi'(v) = (\alpha+1)\frac{-[-\log(v)]^{\alpha}}{v},$$

$$K(v) = v - \frac{v\log(v)}{\alpha+1} \text{ and } \left[\phi^{-1}(v)\right]' = -\frac{v^{\frac{-\alpha}{\alpha+1}}}{\alpha+1}\exp\left(-v^{\frac{1}{\alpha+1}}\right).$$

Hence we have

$$X = F_x^{-1}\left(1 - V^{U^{\frac{1}{\alpha+1}}}\right) \text{ and } Y = F_y^{-1}\left(1 - V^{(1-U)^{\frac{1}{\alpha+1}}}\right).$$

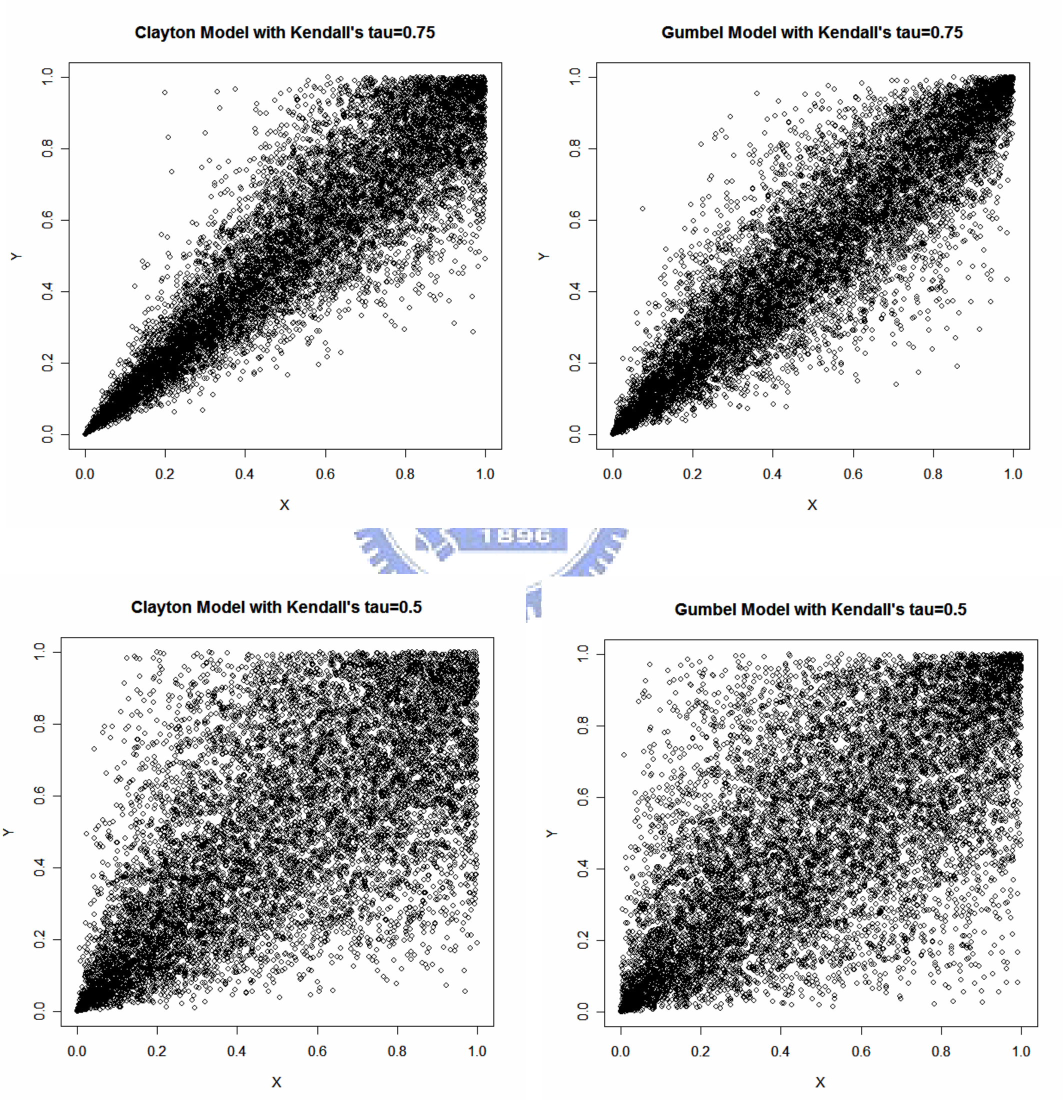
5.4: Comparisons of the Three Approaches

For the frailty approach, to generate a random replication of (X, Y), we need to generate γ and a pair of uniform random variables. For the latter two approaches, we only need to generate a pair of uniform random variables. Hence the frailty approach requires generating at least 50% more random numbers. This is considered as a drawback. For the Clayton model in which γ follows the Gamma distribution, the algorithm is simpler. However for the situation with an arbitrary distribution of γ, to generate a random replicate of γ needed additional work. Moreover, not all of AC family can be derived from frailty model, that is, not every generator $\phi(\cdot)$ can be expressed as an inverse function of Laplace transform of some random variable.

Although the idea of the conditional distribution approach is straightforward, the solution of X_k in (5.2) usually does not have a closed-form expression even for the bivariate case. It is very time consuming if we have to solve the complicated equation numerically.

The proposed method is friendlier compared with the previous two methods. In comparison with the frailty approach, we do not have to generate random numbers, namely γ, which are used only for a temporary purpose. Compared with the conditional distribution approach, our method is technically easier to handle. Sometimes the inverse of $K(\cdot)$ has an explicit form. If not, we can take advantage of the monotone property of $K(\cdot)$ and obtain its inverse using the bisection method. Despite the simplicity of the proposed method, currently the result of Genest and Rivest (1993) can not handle higher dimension with $p > 2$. It implies that we need more general theoretical results in order to extend the proposed algorithm to general multivariate situations.

In Figures 5.1, we plot the generated data using the proposed algorithm. The two models appear to be similar when the level of tau decreases.



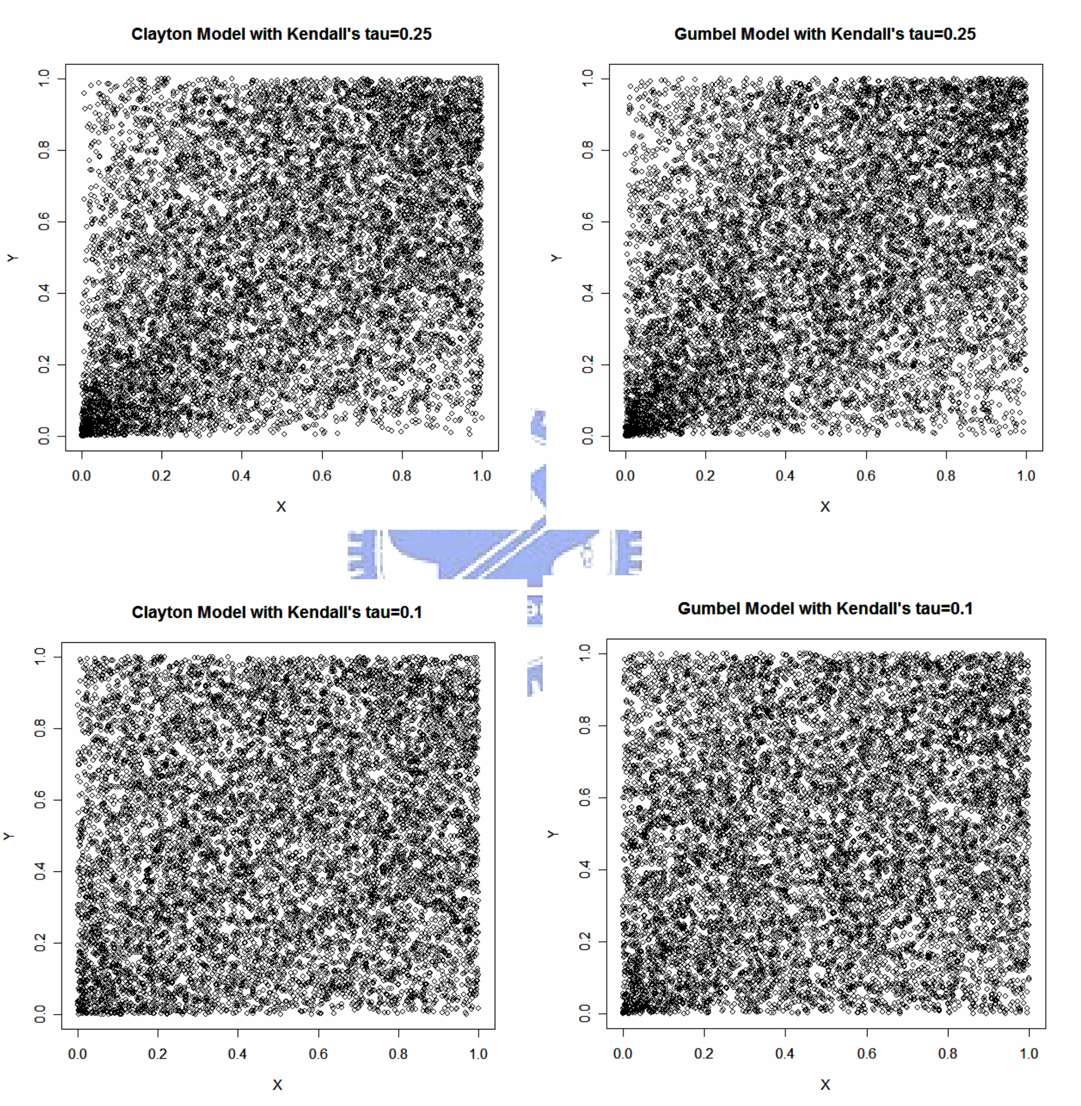


Fig.5.1. Simulated Data using the Proposed Data Generation Algorithm

Chapter 6: Numerical Analysis

Here we examine the performance of the proposed test by simulations. Since we expect our proposed test can be applied to any Archimedean Copula model, we use the Gumbel model for illustration. Recall that for the Gumbel model, we have $\phi(v)=\{-\log(v)\}^{\alpha+1}$ and $\theta_{\alpha}\left(S(x,y)\right)=1-\frac{\alpha}{\log S(x,y)}$.

We generate bivariate failure times following the Gumbel model, also called the positive stable frailty model. We evaluate the performances under different Kendall's τ equal to 0.3, 0.4, 0.5, 0.6 and 0.7 respectively. The marginal distributions of two variables are both exponential with means equal to 1. The bivariate censoring variables are mutually independent and also following exponential distributions such that the probability of censoring is from 0 to 0.5 respectively in each coordinate.

After estimating the association parameter α, we have $\hat{\alpha}$ and $\hat{\alpha}_w$ and let $\hat{\gamma}=\log\hat{\alpha}$ and $\hat{\gamma}_w=\log\hat{\alpha}_w$. Then estimate the variance of $\hat{\gamma}_w-\hat{\gamma}$, $\hat{\sigma}^2_{Jackknife}$. The Gumbel model is rejected if the test statistic

$$T=\frac{\left|\hat{\gamma}_w-\hat{\gamma}\right|}{\hat{\sigma}_{Jackknife}}$$

is greater than $Z_{0.975}=1.96$. In order to assess the power of the proposed test, we also generate the data from other AC models. Based on 100 replications, the empirical probabilities of accepting the Gumbel model under different settings are reported.

Table 6.1 and 6.2 report the empirical probabilities of choosing the Gumbel model. When the true model is Gumbel's, the nominal probability should be 0.95. When the true model is Clayton's or Frank's, the probability is the estimate of type II error rate. Hence we hope that under Gumbel model is correct the proportion of choosing Gumbel should be close to 95/100, and the power is as large as possible. From table 6.1, we find that type-I error is a little smaller than 0.05 when τ equals to 0.3. This may result from the variance estimator using the Jackknife method. The Jackknife algorithm tends to overestimate the variance and results in lower type-I error. When the sample size increases to 200, we see some improvement. Specifically the results in Table 6.2 give more accurate type I probabilities and better power in Table 6.4 and Table 6.6. In Table 6.3 and Table 6.4, we evaluate the type II error probabilities when the true model is Clayton model. In Table 6.5 and Table 6.6, we evaluate the type II error probabilities when the true model is Frank model. From Table 6.3 to Table 6.6, we find that the power deceases as Kendall's τ decreases. This is reasonable, since these three models will all reduce to independent models as Kendall's τ tends to be zero. That is, $\Pr(X > x, Y > y) = S_x(x) \cdot S_y(y)$. This implies that it gets more difficult to distinguish the two models when they are similar.

Figure 6.1 to Figure 6.4 show the powers under true model is Clayton and Frank model with sample size equal to 100 and 200 respectively.

Table 6.1: Empirical Probabilities of Accepting the Gumbel Model with n =100

Censor proportion = 0%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Gumbel				
Sample Mean	-0.038	-0.029	-0.02	0.012	0.047
Sample Standard Deviation	0.88	0.959	1.01	0.993	0.99
Proportion of choosing Gumbel	*99/100*	*97/100*	*96/100*	*96/100*	*95/100*
Censor proportion = 20%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Gumbel				
Sample Mean	-0.115	-0.127	-0.141	-0.133	-0.134
Sample Standard Deviation	0.893	0.968	1.036	1.018	0.987
Proportion of choosing Gumbel	*98/100*	*93/100*	*95/100*	*97/100*	*96/100*
Censor proportion = 50%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Gumbel				
Sample Mean	-0.255	-0.252	-0.234	-0.207	-0.199
Sample Standard Deviation	0.882	0.933	0.941	0.886	0.838
Proportion of choosing Gumbel	*97/100*	*96/100*	*95/100*	*96/100*	*99/100*

Table 6.2: Empirical Probabilities of Accepting the Gumbel Model with n =200

	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 0%	Gumbel				
Sample Mean	0.146	0.16	0.15	0.164	0.132
Sample Standard Deviation	0.986	1.033	1.022	1.006	1.001
Proportion of choosing Gumbel	*97/100*	*95/100*	*92/100*	*93/100*	*93/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 20%	Gumbel				
Sample Mean	0.097	0.109	0.088	0.118	0.114
Sample Standard Deviation	1.031	1.087	1.054	1.021	1.009
Proportion of choosing Gumbel	*95/100*	*94/100*	*94/100*	*93/100*	*96/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 50%	Gumbel				
Sample Mean	-0.012	0.006	-0.034	-0.032	-0.032
Sample Standard Deviation	0.946	0.923	0.856	0.907	0.879
Proportion of choosing Gumbel	*95/100*	*97/100*	*98/100*	*97/100*	*96/100*

Table 6.3:Empirical Type II Error Probabilities of Accepting the Gumbel Model when the True Model is Clayton with n =100

Censor proportion = 0%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Clayton				
Sample Mean	-2.458	-3.203	-3.721	-4.118	-4.358
Sample Standard Deviation	1.113	1.226	1.194	1.193	1.246
Proportion of choosing Gumbel	*30/100*	*17/100*	*6/100*	*3/100*	*3/100*
Censor proportion = 20%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Clayton				
Sample Mean	-1.826	-2.397	-2.793	-3.113	-3.382
Sample Standard Deviation	1.034	1.108	1.108	1.13	1.236
Proportion of choosing Gumbel	*55/100*	*39/100*	*24/100*	*10/100*	*12/100*
Censor proportion = 50%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Clayton				
Sample Mean	-1.031	-1.379	-1.64	-1.919	-2.137
Sample Standard Deviation	0.879	0.983	1.059	1.144	1.135
Proportion of choosing Gumbel	*83/100*	*72/100*	*65/100*	*58/100*	*52/100*

Table 6.4:Empirical Type II Error Probabilities of Accepting the Gumbel Model when the True Model is Clayton with n =200

	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 0%	Clayton				
Sample Mean	-3.644	-4.76	-5.65	-6.303	-6.78
Sample Standard Deviation	1.217	1.511	1.695	1.832	1.894
Proportion of choosing Gumbel	*6/100*	*2/100*	*0/100*	*0/100*	*0/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 20%	Clayton				
Sample Mean	-2.876	-3.782	-4.501	-5.055	-5.432
Sample Standard Deviation	1.083	1.342	1.525	1.672	1.738
Proportion of choosing Gumbel	*23/100*	*6/100*	*3/100*	*0/100*	*0/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 50%	Clayton				
Sample Mean	-1.771	-2.405	-2.882	-3.314	-3.628
Sample Standard Deviation	0.951	1.218	1.392	1.549	1.579
Proportion of choosing Gumbel	*58/100*	*31/100*	*28/100*	*24/100*	*15/100*

Table 6.5:Empirical Type II Error Probabilities of Accepting

the Gumbel Model when the True Model is Frank with n =100

Censor proportion = 0%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Frank				
Sample Mean	-1.865	-2.248	-2.547	-2.807	-2.977
Sample Standard Deviation	0.959	0.949	0.944	0.936	0.964
Proportion of choosing Gumbel	*52/100*	*39/100*	*22/100*	*17/100*	*13/100*
Censor proportion = 20%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Frank				
Sample Mean	-1.666	-2.047	-2.293	-2.552	-2.691
Sample Standard Deviation	0.945	0.957	0.928	0.92	0.956
Proportion of choosing Gumbel	*67/100*	*50/100*	*38/100*	*21/100*	*17/100*
Censor proportion = 50%	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
	Frank				
Sample Mean	-1.24	-1.53	-1.729	-1.961	-2.018
Sample Standard Deviation	0.923	1.03	1.029	1.056	0.999
Proportion of choosing Gumbel	*78/100*	*67/100*	*60/100*	*58/100*	*53/100*

Table 6.6:Empirical Type II Error Probabilities of Accepting the Gumbel Model when the True Model is Frank with n =200

	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 0%	Frank				
Sample Mean	-2.829	-3.426	-3.882	-4.255	-4.597
Sample Standard Deviation	1.194	1.329	1.309	1.264	1.199
Proportion of choosing Gumbel	*21/100*	*12/100*	*4/100*	*2/100*	*1/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 20%	Frank				
Sample Mean	-2.77	-3.382	-3.837	-4.142	-4.363
Sample Standard Deviation	1.167	1.333	1.345	1.346	1.315
Proportion of choosing Gumbel	*24/100*	*15/100*	*5/100*	*2/100*	*2/100*
	tau=0.3	tau=0.4	tau=0.5	tau=0.6	tau=0.7
Censor proportion = 50%	Frank				
Sample Mean	-2.229	-2.762	-3.139	-3.281	-3.413
Sample Standard Deviation	1.176	1.385	1.442	1.376	1.335
Proportion of choosing Gumbel	*39/100*	*30/100*	*24/100*	*21/100*	*18/100*

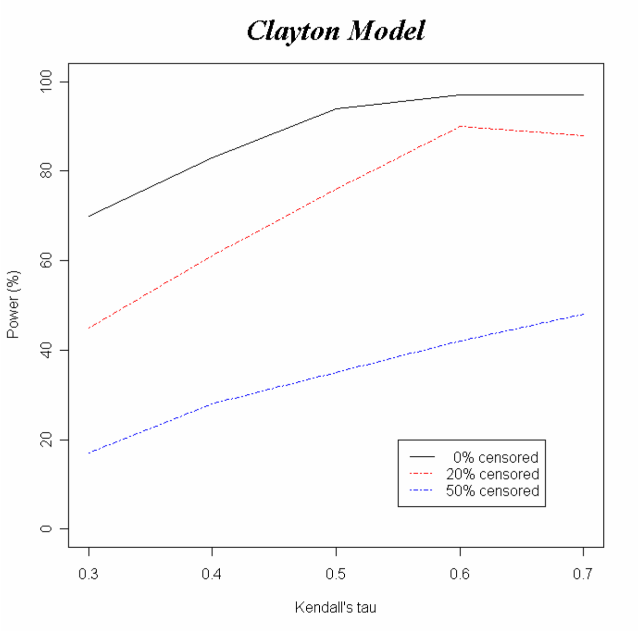


Fig.6.1: Curves of empirical power for H_0: Gumbel vs. H_a: Clayton (n=100)

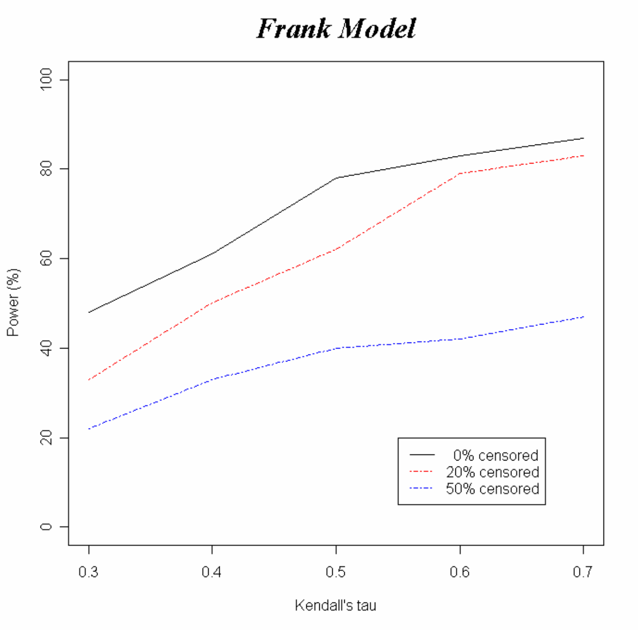


Fig.6.2: Curves of empirical power for H_0: Gumbel vs. H_a: Frank (n=100)

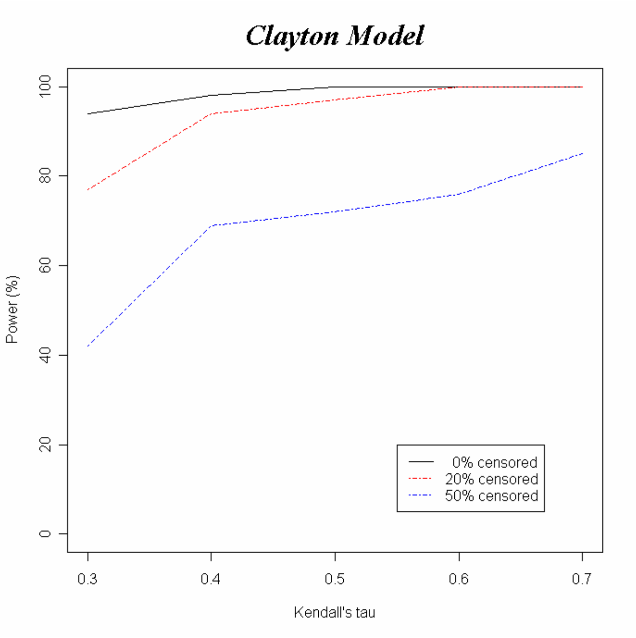


Fig.6.3: Curves of empirical power for H_0: Gumbel vs. H_a: Clayton (n=200)

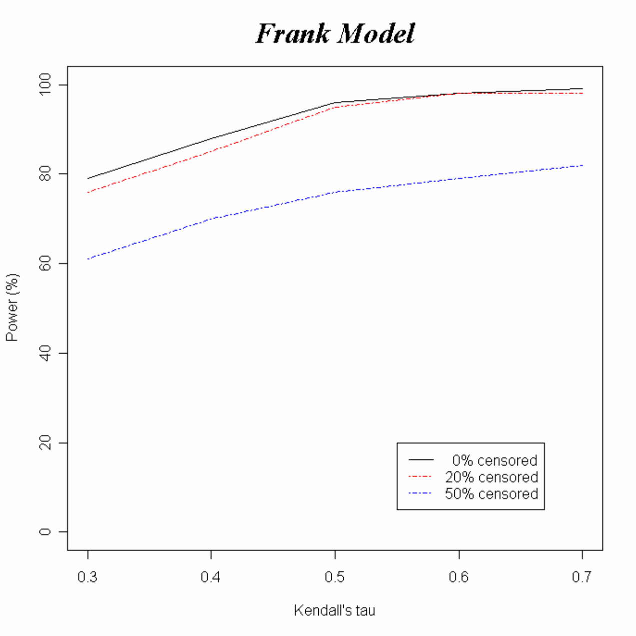


Fig.6.4: Curves of empirical power for H_0: Gumbel vs. H_a: Frank (n=200)

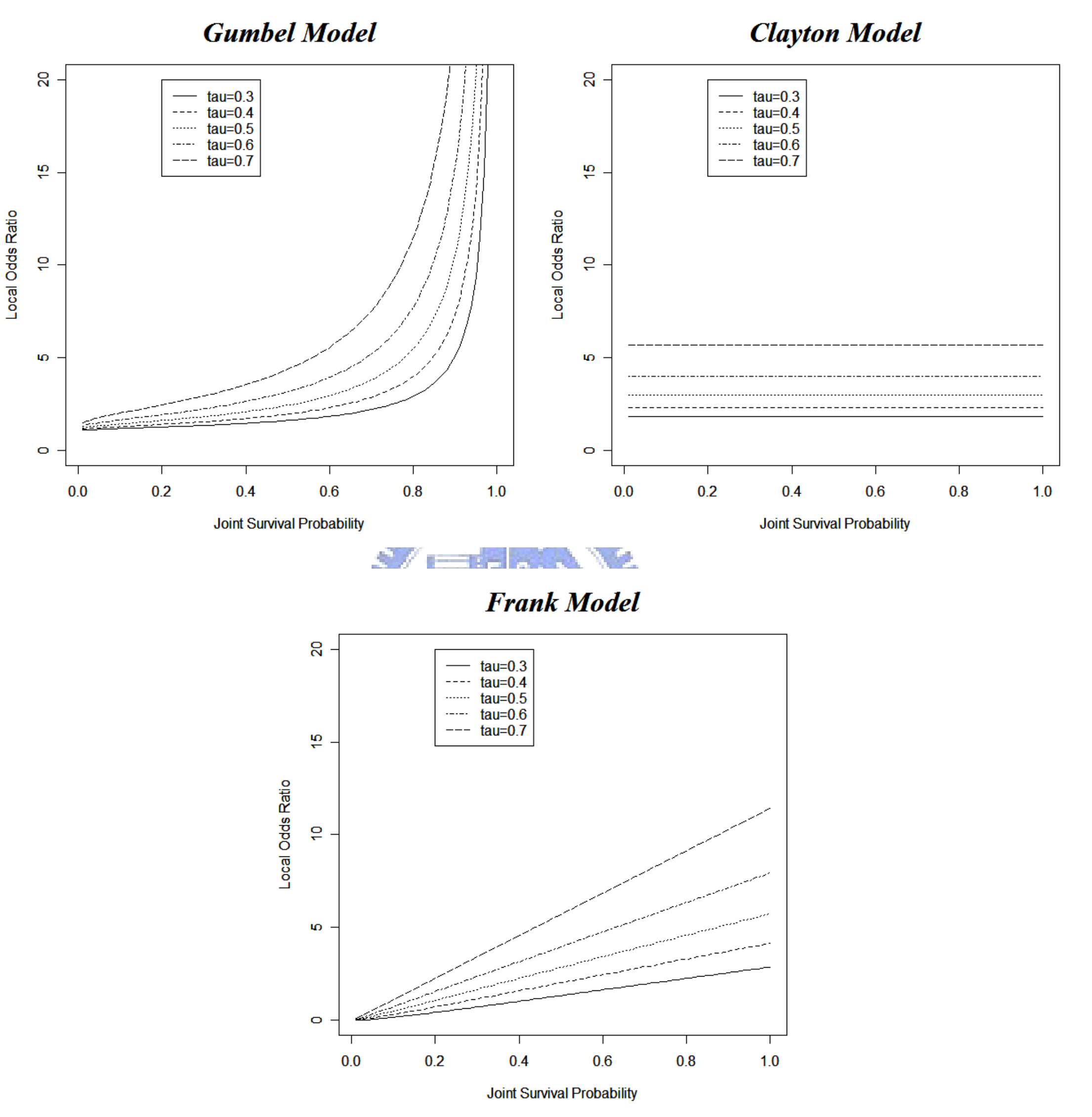


Fig.6.5: The local odds ratio functions at different levels of Kendall's tau for the Gumbel model, the Clayton model and the Frank model

Chapter 7: Conclusion

In this article, we propose a test for checking whether the data following an AC model. In our analysis, we use the Gumbel model for illustration. To verify whether proposed test statistic is asymptotically normal, and we examine its distribution by simulations. Our conjecture is confirmed. We have also found that the power of the proposed test is satisfactory. Shih (1998) has analyzed the situation when the null hypothesis is the Clayton model while the alternative hypothesis is Gumbel's model. In our simulations, we reverse the roles of the two models in setting the hypotheses. Our result is similar to that of Shih.

The power decreases as the censoring proportion increases. When the null hypothesis is the Gumbel model, the power is higher under the Clayton alternative than under the Frank model. Recall that in Figure 6.5., the Gumbel model is more close to the Frank model and less similar to the Clayton model. It is easier to distinguish two models which are more different which results in higher power.

As for future investigation, we may try more model combinations. Also it may be interesting to compare the proposed test with the test of Wang and Wells (2000) by simulations.

Appendix

Here, we prove the survival version of the theorem in Genest & Rivest. The proof can be divided into several parts.

Consider the survival AC model:

$X,Y \sim C(1-x,1-y)=\phi^{-1}\{\phi(1-x)+\phi(1-y)\}=\Pr(X>x,Y>y)$,

$C(1-x,1)=1-x$, $C(1,1-y)=1-y$

Define the transformation:

$U=\phi(1-X)/\{\phi(1-X)+\phi(1-Y)\}, V=\phi^{-1}\{\phi(1-X)+\phi(1-Y)\}$

We show that (a) $U \sim Unif(0,1)$, (b) V has c.d.f $K(v)=v-\phi(v)/\phi'(v)$,

(c) $U \perp V$.

(i) Define $S=\phi(1-X)$ and $T=\phi(1-Y)$. Show that the joint survival function of (S,T) can be written as $\phi^{-1}(s+t)$.

(ii) Show the formula

$$\frac{d}{dt}\phi^{-1}(t)=\frac{1}{\phi'\left(\phi^{-1}(t)\right)}$$

(iii) Show that the conditional survival function can be written as

$$\Pr(S>s\,|\,T=t)=\frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'\left[\phi^{-1}(s+t)\right]}$$

(iv) Show the relation

$$\Pr(U>u,V>v)=\int_0^\infty \Pr\left\{\frac{ut}{1-u}<S<\phi(v)-t\,|\,T=t\right\}\Pr(T=t)$$

(v) Obtain $\Pr(U>u,V>v)\ =(1-u)\cdot\left(1+\frac{\phi(v)}{\phi'(v)}-v\right)$

(i)

$$\begin{aligned}\Pr(S>s,T>t)&=\Pr\left(\phi(1-X)>s,\phi(1-Y)>t\right)\\&=\Pr\left(1-X<\phi^{-1}(s),1-Y<\phi^{-1}(t)\right)\\&=\Pr\left(X>1-\phi^{-1}(s),Y>1-\phi^{-1}(s)\right)\\&=\phi^{-1}\left\{\phi\left(\phi^{-1}(s)\right)+\phi\left(\phi^{-1}(t)\right)\right\}\\&=\phi^{-1}(s+t)\end{aligned}$$

(ii)

$$\because \phi\left(\phi^{-1}(t)\right)=t$$

$$\frac{d\phi\left(\phi^{-1}(t)\right)}{dt}=1$$

$$\Rightarrow \phi'\left(\phi^{-1}(t)\right)\cdot\frac{d\phi^{-1}(t)}{dt}=1$$

$$\therefore \frac{d\phi^{-1}(t)}{dt}=\frac{1}{\phi'\left(\phi^{-1}(t)\right)}$$

(iii)

$$\begin{aligned}\Pr(T\le t)&=\Pr\left(\phi(1-Y)\le t\right)\\&=\Pr\left(1-Y\ge\phi^{-1}(t)\right)\\&=\Pr\left(Y\le 1-\phi^{-1}(t)\right)\\&=1-\phi^{-1}(t)\quad\left(,\text{since } Y\sim U(0,1)\right)\end{aligned}$$

$$\Pr(S>s\mid T=t)=\frac{\Pr(S>s,T=t)}{\Pr(T=t)}$$

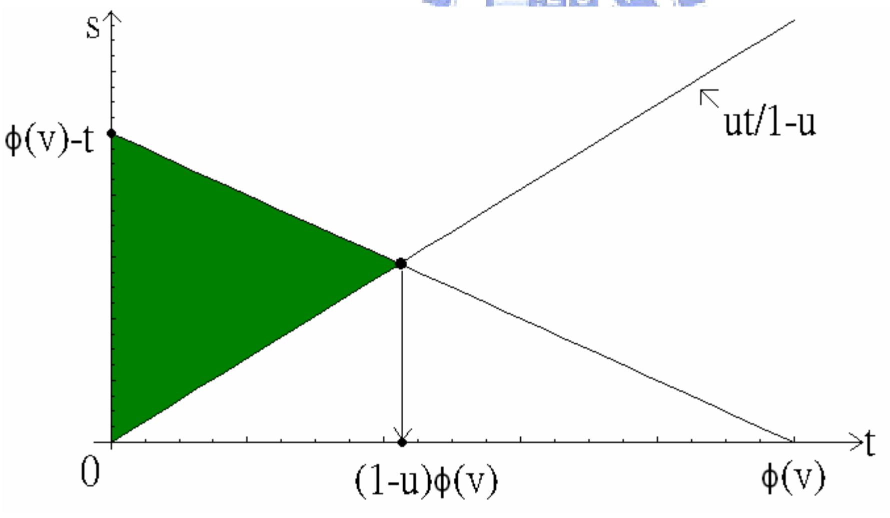
$$\begin{aligned}\frac{\Pr(S>s,T=t)}{dt}&=-\frac{\partial}{\partial t}\Pr(S>s,T>t)\\&=-\frac{\partial}{\partial t}\phi^{-1}(s+t)\\&=\frac{-1}{\phi'\left(\phi^{-1}(s+t)\right)}\end{aligned}$$

$$
\begin{aligned}
\frac{\Pr(T=t)}{dt} &= \frac{\partial}{\partial t}\Pr(T\le t)\\
&= \frac{-1}{\phi'\left[\phi^{-1}(t)\right]}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(S>s\,|\,T=t) &= \frac{\Pr(S>s,T=t)}{\Pr(T=t)}\\
&= \frac{\Pr(S>s,T=t)/dt}{\Pr(T=t)/dt}\\
&= \frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'\left[\phi^{-1}(s+t)\right]}
\end{aligned}
$$

(iv)&(v)

$$
\begin{aligned}
\Pr(U>u,V>v) &= \Pr\left(S>\frac{uT}{1-u},S<\phi(v)-T\right)\\
&= \int_{t=0}^{\infty}\Pr\left(\frac{uT}{1-u}<S<\phi(v)-T\,|\,T=t\right)dF(t)\\
&= \int_{t=0}^{(1-u)\phi(v)}\Pr\left(\frac{ut}{1-u}<S<\phi(v)-t\,|\,T=t\right)dF(t)
\end{aligned}
$$



$$\left(\phi(v)-t=\frac{ut}{1-u}\Rightarrow t=(1-u)\phi(v)\right)$$

$$\because \Pr\left(S\le\frac{ut}{1-u}\,|\,T=t\right)=1-\frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'\left[\phi^{-1}\left(\frac{t}{1-u}\right)\right]},\text{ and }\Pr\left(S\le\phi(v)-t\,|\,T=t\right)=1-\frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'(v)}.$$

$$\therefore \Pr(U>u,V>v)$$

$$=\int_{t=0}^{(1-u)\phi(v)}\left[\Pr\left(S\le\phi(v)-t\,|\,T=t\right)-\Pr\left(S\le\frac{ut}{1-u}\,|\,T=t\right)\right]dF(t)$$

$$=\int_{t=0}^{(1-u)\phi(v)}\left[\frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'\left[\phi^{-1}\left(\frac{t}{1-u}\right)\right]}-\frac{\phi'\left[\phi^{-1}(t)\right]}{\phi'(v)}\right]\cdot\frac{-1}{\phi'\left[\phi^{-1}(t)\right]}dt$$

$$=\int_{t=0}^{(1-u)\phi(v)}\left[\frac{1}{\phi'(v)}-\frac{1}{\phi'\left[\phi^{-1}\left(\frac{t}{1-u}\right)\right]}\right]dt$$

$$=\frac{t}{\phi'(v)}-(1-u)\cdot\phi^{-1}\left(\frac{t}{1-u}\right)\Bigg|_{t=0}^{(1-u)\phi(v)}$$

$$=(1-u)\cdot\frac{\phi(v)}{\phi'(v)}-(1-u)\cdot v+(1-u)$$

$$=(1-u)\cdot\left(1+\frac{\phi(v)}{\phi'(v)}-v\right)$$

$$\begin{aligned}\therefore F_U(u)&=1-(1-u)\\&=u\end{aligned}$$

$$\begin{aligned}F_V(v)&=1-\left(1+\frac{\phi(v)}{\phi'(v)}-v\right)\\&=v-\frac{\phi(v)}{\phi'(v)}\\&=K(v)\end{aligned}$$

and $U\perp V$.

Here, we try to prove the asymptotic normality of $\hat{\gamma}_w - \hat{\gamma}$. The idea is that, first prove the asymptotic normality of untransformed estimator $\hat{\alpha}_W - \hat{\alpha}_{Uw}$, then utilize delta method to derive the asymptotic normality of $\hat{\gamma}_w - \hat{\gamma}$.

$$\sqrt{n}\left(\hat{\alpha}_{W.Dab} - \hat{\alpha}_{Uw.Dab}\right) \longrightarrow N\left(0, \sigma^2\right)\sigma^2$$

$$= \sqrt{n}\left(\hat{\alpha}_{W.Dab} - \alpha\right) - \sqrt{n}\left(\hat{\alpha}_{Uw.Dab} - \alpha\right)$$

$$= \sqrt{n}\left(\hat{\alpha}_{W.Dab} - \hat{\alpha}_{W.true}\right) \xrightarrow{p} 0$$

$$+\sqrt{n}\left(\hat{\alpha}_{W.true} - \alpha\right) \xrightarrow{d} N\left(0, \sigma_W^2\right)$$

$$-\sqrt{n}\left(\hat{\alpha}_{Uw.Dab} - \hat{\alpha}_{Uw.true}\right) \xrightarrow{p} 0$$

$$+\sqrt{n}\left(\hat{\alpha}_{Uw.true} - \alpha\right) \xrightarrow{d} N\left(0, \sigma_{Uw}^2\right)$$

$$S(\alpha) = \sum_{i<j} \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}{\left[\left(R_{ij} - 1\right) + \theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)\right]} \cdot \frac{d \ln \theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{d\alpha} \left\{\Delta_{ij} - \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}\right\}$$

$$= n^{-1} \sum_{i<j} \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}{\left[\frac{\left(R_{ij} - 1\right)}{n} + \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{n}\right]} \cdot \frac{d \ln \theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{d\alpha} \left\{\Delta_{ij} - \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}\right\}$$

$$R_{ij} = R\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) = \sum_{i=1}^{n} I\left(X_i \geq \tilde{X}_{ij}, Y_i \geq \tilde{Y}_{ij}\right)$$

$$\pi(x, y) = \Pr\left(X > x, Y > y\right)$$

$\frac{R_{ij}}{n} \xrightarrow{p} \pi\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)$ by SLLN, and $\frac{1}{n} \longrightarrow 0$

We suppose that $\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)$ is bounded, then

$$\frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{n} \longrightarrow 0$$

$$S(\alpha) \approx n^{-1} \sum_{i<j} \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}{\pi\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)} \cdot \frac{\theta'_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)} \left\{\Delta_{ij} - \frac{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right)}{\theta_\alpha\left(\tilde{X}_{ij}, \tilde{Y}_{ij}\right) + 1}\right\}$$

$$= n^{-1}\sum_{i<j} h_\alpha\left[\begin{pmatrix}X_i\\ \delta_i^X\\ Y_i\\ \delta_i^Y\end{pmatrix},\begin{pmatrix}X_j\\ \delta_j^X\\ Y_j\\ \delta_j^Y\end{pmatrix}\right]$$

Here noticed that $\theta_\alpha\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)$, $\pi\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)$, $\theta'_\alpha\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)$ and Δ_{ij} can be obtained only by (i, j) pairs observations. So, we can utilize the U -statistic to derive the analytic properties of $S(\alpha)$.

Reference:

CLAYTON, D. G. (1978). A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. **65**, 141-51.

DABROWSKA, D. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics*. **16**, 1475-89.

FREES EW, VALDEZ E. (1998). Understanding the relationships using copulas . *North American Actuarial Journal*. **2**, 1-25.

GENEST, C. & RIVEST, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*. **88**, 1034-43.

LEE, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician*. **47**, 209-215.

OAKES, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*. **73**, 353-61.

OAKES, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*. **84**, 487-93.

SHIH, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika* **85**, 189-200.

WANG, W. & WELLS, M. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika*. **84**, 863-880.