Statistica Sinica **20** (2010), 441-453

GOODNESS-OF-FIT TESTS FOR ARCHIMEDEAN COPULA MODELS

Antai Wang

Georgetown University Medical Center

Abstract: In this paper, we propose two tests for parametric models belonging to the Archimedean copula family, one for uncensored bivariate data and the other one for right-censored bivariate data. Our test procedures are based on the Fisher transform of the correlation coefficient of a bivariate (U, V), which is a one-to-one transform of the original random pair (T_1, T_2) that can be modeled by an Archimedean copula model. A multiple imputation technique is applied to establish our test for censored data and its p value is computed by combining test statistics obtained from multiply imputed data sets. Simulation studies suggest that both procedures perform well when the sample size is large. The test for censored data is carried out for a medical data example.

Key words and phrases: Archimedean copula models, bivariate survival data, the Fisher transform, the Kendall distribution.

1. Introduction

The problem of specifying a probability model for independent observations $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ from a bivariate population with continuous survivor function $S(t_1, t_2)$ can be simplified by expressing S in terms of its marginals, $S_1(t_1)$, $S_2(t_2)$, and their associated dependence function, C, implicitly defined through the identity

$$S(t_1, t_2) = C\Big\{S_1(t_1), S_2(t_2)\Big\}.$$

Here C, called a copula, characterizes the dependence structure between the random variables T_1, T_2 (see Joe (1997) and Nelsen (1999)). Because of its simple form, copula models have been widely used to model multivariate data.

A bivariate survivor function $S(t_1, t_2)$ with marginal survivor functions $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$ is defined to be generated by an "Archimedean copula" if it can be expressed in the form $S(t_1, t_2) = q^{-1}[q\{S_1(t_1)\} + q\{S_2(t_2)\}]$ for some convex, decreasing function q defined on $[0, 1]$ with $q(1) = 0$ (see Genest and Rivest (1993) and Genest and MacKay (1986)). A large class of Archimedean copulas arise naturally from bivariate frailty models (Oakes (1989) and Nelsen (1999)) in which T_1 and T_2 are conditionally independent given an unobserved 'frailty' W (here W is common to both T_1 and T_2) and each follows proportional hazards model in W. Denote the distribution function of W as $F(\cdot)$. Take $p(\cdot) = q^{-1}(\cdot)$ where $p(s) = E(e^{-sW})$, the Laplace transform of the $F(\cdot)$. The first model of such a type in the survival analysis literature was proposed by Clayton (1978). He used the Gamma distribution with $p(s) = (1+s)^{-1/\alpha}$, which leads to the bivariate survivor function

$$S(t_1, t_2) = \left\{ \frac{1}{S_1(t_1)^{-\alpha} + S_2(t_2)^{-\alpha} - 1} \right\}^{1/\alpha},$$

for $\alpha > 0$, see Oakes (1982). Another important frailty model, the Frank model (Clayton (1978)), has $p(s) = -\log\{1 - (1 - e^{-\beta})/e^s\}/\beta$; its bivariate survivor function is

$$S(t_1, t_2) = -\frac{1}{\beta} \log \left[\frac{\{\exp(-\beta) - 1 + (\exp\{-\beta S_1(t_1)\} - 1)(\exp\{-\beta S_2(t_2)\} - 1)\}}{(\exp(-\beta) - 1)} \right]$$

for $\beta \neq 0$. Other models, such as the Hougaard model (Hougaard (1986)) and the Inverse Gaussian copula model, belong to this family.

Many authors have proposed goodness-of-fit tests for models belonging to the Archimedean copula family. Oakes (1989) proposed a graphic diagnostic approach to check the goodness-of-fit for such type of models. Shih (1998) proposed a goodness-of-fit test for the Clayton model. That can be applied to both uncensored and censored data. However, the test procedure is designed specifically for the Clayton model. Wang and Wells (2000a) proposed a model selection procedure within the Archimedean copula family for right-censored bivariate data based on the so-called L^2 norm of the Kendall distribution (basically a distance measure between the empirical and the estimated Kendall distribution). Genest, Quessy, and Rémillard (2006) extended the idea in Wang and Wells (2000a) and proposed a general goodness-of-fit test procedure for models belonging to the Archimedean copula family. Their test is designed for uncensored data and it does not seem to allow an easy extension to right-censored bivariate data. Anderson, Ekstraum, Klein, Shu and Zhang (2005) proposed a bootstrap goodness-of-fit test for copula models that can be applied to bivariate survival data. The bootstrap p-value of their test is calculated based on a procedure that involves generating both the original bivariate data and the independent censoring data. It turns out that none of above test statistics has a simple distribution under the null hypothesis, and most of them are quite computationally intensive. For censored bivariate data, there has not been any simple goodness-of-fit test for Archimedean copula models.

In this paper, we propose a simple goodness-of-fit test to check the Archimedean copula model assumption for uncensored bivariate data. We then extend our test to right-censored data by combining p values obtained from a multiple imputation procedure. Simulation studies suggest that our test procedures work quite well when the sample size is large.

Our paper is organized in the following way. In Section 2, we propose our test for uncensored data. We then extend our test to censored bivariate data in Section 3. Simulation studies are presented in Section 4. The test for censored data is used on the DRS data in Section 5. We end our paper with some discussion in Section 6.

2. A Simple Test for Uncensored Data

Genest and Rivest (1993) have shown that if (T_1, T_2) follows an Archimedean copula with the marginal survivor functions $S_1(t_1)$ and $S_2(t_2)$, then

$$U = \frac{q(S_1(T_1))}{q\{S(T_1,T_2)\}}, \quad V = S(T_1,T_2) = p\Big[q\{S_1(T_1)\} + q\{S_2(T_2)\}\Big]$$

are independently distributed random variables (here $p(\cdot) = q^{-1}(\cdot)$), U is $[0,1]$ and V follows the so-called "the Kendall distribution" with the density function

$$k(v) = \frac{q(v)q''(v)}{q'(v)^2}$$

defined on $[0,1]$, where q as a function of v depends on an unknown parameter θ. Our null hypothesis is that (T_1, T_2) follows some Archimedean copula model C_θ. Under this assumption, we know that the corresponding random variables U and V are independent, which means that their correlation coefficient $\rho = E[U - E(U))(V - E(V))]/\sqrt{\text{var}(U)\text{var}(V)} = 0$. Hence a goodness-of-fit test for $H_0 : C = C_\theta$ where C_θ is a parametric model belonging to the Archimedean copula family versus $H_1 : C \neq C_\theta$ can be constructed based on a test procedure for the null hypothesis $H_0' : \rho = 0$ versus $H_1' : \rho \neq 0$.

In reality, we cannot observe U and V. However, they can be consistently estimated by

$$\hat{U} = \frac{q_{\hat{\theta}}\{\hat{S}_1(T_1)\}}{q_{\hat{\theta}}\{\hat{S}(T_1,T_2)\}} \quad \text{and} \quad \hat{V} = \hat{S}(T_1,T_2)$$

respectively, where $\hat{S}_1(T_1)$, $\hat{S}(T_1,T_2)$ are empirical marginal and joint survivor functions, and $\hat{\theta}$ is a consistent estimator of the unknown parameter θ. Such an estimator can be the moment estimator proposed by Genest and Rivest (1993) or the pseudo-MLE proposed by Genest, Ghoudi and Rivest (1995). In the case where margins are estimated within a parametric model, a two-stage estimation procedure (Shih and Louis (1995)) can be applied. Basically, we first estimate the unknown parameter in the marginal distributions, then we estimate the dependence parameter θ by solving the score equation based on the pseudo-observations $(\hat{S}_1(T_{1i}), \hat{S}_2(T_{2i}))$ for $i \in \{1, \ldots, n\}$. Under suitable regularity conditions, consistency of above estimators has been established.

By the Continuous Mapping Theorem, $(\hat{U}, \hat{V}) \to (U, V)$ in distribution when $n \to \infty$. Therefore, when the sample size is large, the distribution of $(\hat{U}, \hat{V})$ is approximately that of (U, V) with joint density function $k(v)$ defined on $[0,1] \times [0,1]$ (since U and V are independent with U being a Uniform[0, 1] and V having the Kendall density $k(v)$ defined above).

Let

$$r_n = \frac{\sum_{i=1}^n (\hat{U}_i - \bar{\hat{U}})(\hat{V}_i - \bar{\hat{V}})}{\sqrt{\sum_{i=1}^n (\hat{U}_i - \bar{\hat{U}})^2 \sum_{i=1}^n (\hat{V}_i - \bar{\hat{V}})^2}},$$

where

$$\hat{U}_i = \frac{q_{\hat{\theta}}\{\hat{S}_1(T_{1i})\}}{q_{\hat{\theta}}\{\hat{S}(T_{1i}, T_{2i})\}}, \quad \hat{V}_i = \hat{S}(T_{1i}, T_{2i}),$$

and $\bar{\hat{U}}$, $\bar{\hat{V}}$ are the sample means of $\hat{U}_i$ and $\hat{V}_i$, respectively. Take $Z_n = 1/2 \log\{(1+r_n)/(1-r_n)\}$. Hawkins (1989) proved the following about the asymptotic distribution of Fisher's Z statistic.

Theorem 1. *Suppose that (U_i, V_i), for $i \in \{1 \ldots n\}$, are independently identically distributed random pairs with mean $(0,0)$ and variance $(1,1)$ that follow some bivariate distribution F with finite fourth moments. Then as $n \to \infty$, $\sqrt{n}[Z_n - 1/2 \ln((1+\rho)/(1-\rho))] \to N(0, \tau_F^2)$ in distribution, where $\tau_F^2 = (1-\rho^2)^{-2} 1/4\{(m_{40} + 2m_{22} + m_{04})\rho^2 - 4(m_{31} + m_{13})\rho + 4m_{22}\}$, with $m_{rs} = E_F(U_i^r V_i^s)$ for $r, s \in \{0, 1, 2, 3, 4\}$.*

Under H_0, U and V are independent, so $\rho = 0$ for $\{U - E(U)\}/\sqrt{\text{var}(U)}$ and $\{V - E(V)\}/\sqrt{var(V)}$. Also because of the independence of U and V, we can see that

$$m_{22} = \frac{E[\{U - E(U)\}^2 \{V - E(V)\}^2]}{\text{var}[\{U - E(U)\}]\text{var}[\{V - E(V)\}]} = 1.$$

Noticing the fact that Z_n is location and scale invariant such that Z_n is unchanged after U and V have been replaced by $\{U - E(U)\}/\sqrt{\text{var}(U)}$ and $\{V - E(V)\}/\sqrt{\text{var}(V}$ respectively, we can reach the following conclusion after applying Theorem 1.

Theorem 2. *Under the null hypothesis that (T_{1i}, T_{2i}) for $i \in \{1 \ldots n\}$ are independently identically distributed random pairs that follow some Archimedean copula model C_θ, $\sqrt{n} Z_n \to N(0, 1)$ in distribution.*

A simple test of H_0 for uncensored bivariate data can therefore be established as: reject H_0 at 5% significance level if $|\sqrt{n} Z_n| > Z_{0.975} = 1.96$.

3. A Test for Censored Data

In this section, we propose a test procedure to check the parametric model assumption for models belonging to the Archimedean copula family when bivariate data is subject to right-censoring. We assume that (T_{1i}, T_{2i}), $i = 1, \ldots, n$, are independently identically distributed random pairs whose distribution can be modeled by a specific Archimedean copula. We also assume that (T_{1i}, T_{2i}) are subject to independent right-censoring by censoring vectors (C_{1i}, C_{2i}) for $i = 1, \ldots, n$. Because of the right-censoring, we can only observe $\{(X_{1i}, X_{2i}), (\delta_{1i}, \delta_{2i})\}$ where $X_{1i} = \min\{T_{1i}, C_{1i}\}$, $X_{2i} = \min\{T_{2i}, C_{2i}\}$, $\delta_{1i} = \mathrm{I}\{T_{1i} \le C_{1i}\}$, and $\delta_{2i} = \mathrm{I}\{T_{2i} \le C_{2i}\}$. As a result, we have four different censoring patterns: (1) $\delta_{1i} = \delta_{2i} = 0$, i.e., $T_{1i} > C_{1i} = c_{1i}$ and $T_{2i} > C_{2i} = c_{2i}$; (2) $\delta_{1i} = 1$, $\delta_{2i} = 0$; (3) $\delta_{1i} = 0$, $\delta_{2i} = 1$; (4) $\delta_{1i} = 1$, $\delta_{2i} = 1$. We apply a multiple imputation (MI) procedure to recover the pseudo complete data $(\hat{U}_i, \hat{V}_i)$ for $i = 1, \ldots, n$ (they can not be consistently estimated as before because of the right-censoring) based on different censoring patterns and then establish our test based on multiply imputed complete data sets. The following Theorems are needed, proofs are given in a supplementary file of this paper posted on `http://www.stat.sinica.edu.tw/statistica`.

Theorem 3. *Let (T_1, T_2) have a distribution that can be modelled by an absolutely continuous Archimedean copula. If (T_1, T_2) is subject to independent right-censoring by a censoring vector (C_1, C_2) that follows a bivariate continuous distribution, then we have:*

1. *the distribution function of $\{(U, V)|T_1 > c_1, T_2 > c_2\}$ (i.e., $T_1 > C_1 = c_1$, $T_2 > C_2 = c_2$) is*

$$H_1(u,v)=\begin{cases} \dfrac{[v-\frac{q(v)-q\{S(c_1,c_2)\}}{q'(v)}-(1-u)p\{\frac{q(S_2(c_2))}{1-u}\}]}{S(c_1,c_2)}, & 1-\dfrac{q\{S_2(c_2)\}}{q(v)} < u \le 1 \\ \dfrac{\{uv+\frac{q\{S_1(c_1)\}-uq(v)}{q'(v)}\}}{S(c_1,c_2)}, & \dfrac{q\{S_1(c_1)\}}{q(v)} < u \le 1-\dfrac{q\{S_2(c_2)\}}{q(v)} \\ \dfrac{\{up\{\frac{q\{S_1(c_1)\}}{u}\}\}}{S(c_1,c_2)}, & 0 \le u \le \dfrac{q\{S_1(c_1)\}}{q(v)} \end{cases}$$

for $0 \le v \le S(c_1, c_2)$;

2. *the distribution function of $\{(U, V)|T_1 = t_1, T_2 > c_2\}$ (i.e., $C_1 > T_1 = t_1$, $T_2 > C_2 = c_2$) is*

$$H_2(u,v) = \begin{cases} \dfrac{p'\{q(v)\}}{p'\{q(S(t_1,c_2)\}}, & \dfrac{q\{S_1(t_1)\}}{q(v)} < u \le 1 \\ \dfrac{p'[q\{S_1(t_1)\}/u]}{p'\{q(S(t_1,c_2)\}}, & 0 \le u \le \dfrac{q\{S_1(t_1)\}}{q(v)} \end{cases}$$

for $0 \le v \le S(t_1, c_2)$;

3. *the distribution function of* $\{(U,V)|T_1 > c_1, T_2 = t_2\}$ *(i.e.,* $T_1 > C_1 = c_1$, $C_2 > T_2 = t_2$*) is*

$$H_3(u,v) = \frac{\Big\{p'(q(v)) - p'[q\{S_2(t_2)\}/(1-u)]\Big\}}{p'\{q(S(c_1,t_2)\}}, \quad 1 - \frac{q\{S_2(t_2)\}}{q(v)} \le u \le 1$$

for $0 \le v \le S(c_1,t_2)$.

Corollary 1. *Let* (T_1,T_2) *have a distribution that can be modelled by an absolutely continuous Archimedean copula. If* (T_1,T_2) *is subject to independent right-censoring by a censoring vector* (C_1,C_2) *that follows a bivariate continuous distribution, then we have:*

1. *the distribution function of* $(V|T_1 > c_1, T_2 > c_2)$ *(i.e.,* $T_1 > C_1 = c_1$, $T_2 > C_2 = c_2$*) is*

$$F_1(v,c_1,c_2) = \frac{1}{S(c_1,c_2)}\left[v - \frac{q(v) - q\{S(c_1,c_2)\}}{q'(v)}\right], \quad 0 \le v \le S(c_1,c_2);$$

2. *the distribution function of* $(V|T_1 = t_1, T_2 > c_2)$ *(i.e.,* $C_1 > T_1 = t_1$, $T_2 > C_2 = c_2$*) is*

$$F_2(v,t_1,c_2) = \frac{p'\{q(v)\}}{p'\{q(S(t_1,c_2))\}}, \quad 0 \le v \le S(t_1,c_2);$$

3. *the distribution function of* $(V|T_1 > c_1, T_2 = t_2)$ *(i.e.,* $T_1 > C_1 = c_1$, $C_2 > T_2 = t_2$*) is*

$$F_3(v,c_1,t_2) = \frac{p'\{q(v)\}}{p'\{q(S(c_1,t_2))\}}, \quad 0 \le v \le S(c_1,t_2).$$

Corollary 2. *Let* (T_1,T_2) *have a distribution that can be modelled by an absolutely continuous Archimedean copula* C_θ *with the copula generator* q*. If* (T_1,T_2) *is subject to independent right-censoring by a censoring vector* (C_1,C_2) *that follows a bivariate continuous distribution, then we have;*

1. *the distribution function of* $(U|T_1 > c_1, T_2 > c_2)$ *(i.e.,* $T_1 > C_1 = c_1$, $T_2 > C_2 = c_2$*) is*

$$G_1(u,c_1,c_2) = \begin{cases} 1 - \frac{1-u}{S(c_1,c_2)} p\Big(\frac{q(S_2(c_2))}{1-u}\Big), & \frac{q(S_1(c_1))}{q(S(c_1,c_2))} \le u \le 1 \\ \frac{u}{S(c_1,c_2)} p\Big(\frac{q(S_1(c_1))}{u}\Big) & , \quad 0 \le u \le \frac{q(S_1(c_1))}{q(S(c_1,c_2))}; \end{cases}$$

2. *the distribution function of* $(U|T_1 = t_1, T_2 > c_2)$ *is*

$$G_2(u, t_1, c_2) = \frac{p'\{q(S_1(t_1))/u\}}{p'\{q(S(t_1, c_2)\}}, \quad 0 \leq u \leq \frac{q\{S_1(t_1)\}}{q\{S(t_1, c_2)\}};$$

3. *the distribution function of* $(U|T_1 > c_1, T_2 = t_2)$ *is*

$$G_3(u, c_1, t_2) = 1 - \frac{p'\{q(S_2(t_2))/(1-u)\}}{p'\{q(S(c_1, t_2))\}}, \quad \frac{q\{S_1(c_1)\}}{q\{S(c_1, t_2)\}} \leq u \leq 1.$$

To impute the unknown random vector (U, V) from doubly censored data, we need another result about the conditional distribution of the random variable $(U|V = v, T_1 > c_1, T_2 > c_2)$.

Theorem 4. *Let* $(T_1,\ T_2)$ *have a distribution that can be modeled by an Archimedean copula. The conditional distribution of the random variable* $(U|V = v, T_1 > c_1, T_2 > c_2)$ *is uniformly distributed on the interval* $[q\{S_1(c_1)\}/q(v), 1 - q\{S_2(c_2)\}/q(v)]$.

Remark. when $c_1 = c_2 = 0$, $q\{S_1(c_1)\} = q\{S_2(c_2)\} = q(1) = 0$, the conditional distribution of $(U|V = v, T_1 > c_1, T_2 > c_2)$ is uniform distribution on $[0, 1]$ and independent of V. This result corresponds to the uncensored case and coincides with the fact proved by Genest and Rivest (1993).

Based on previous results, if the original data pairs (T_{1i}, T_{2i}) for $i \in \{1 \ldots n\}$ are subject to independent right-censoring by random pairs (C_{1i}, C_{2i}) for $i \in \{1 \ldots n\}$, our goodness-of-fit test for censored bivariate data can be set up as follows.

1. Estimation step. Estimate the dependence parameter by a semiparametric estimator $\hat{\theta}$ such as the one proposed by Shih and Louis (1995), or the nonparametric estimator proposed by Brown, Hollander and Korwar (1974) for censored data. Estimate the marginal survivor functions by Kaplan-Meier estimates, and the joint survivor function by the Dabrowska estimator (Dabrowska (1988)). For simplicity, we use $(\hat{U}_i, \hat{V}_i)$ to denote the imputed or estimated bivariate random vector corresponding to (U_i, V_i).
2. Data imputation step. For imputation, we replace the unknown parameter and survivor functions with the corresponding estimators obtained from the estimation step.
   (a) When (T_{1i}, T_{2i}) is doubly censored, i.e., $T_{1i} > c_{1i}$, $T_{2i} > c_{2i}$ for some observed pair $(C_{1i}, C_{2i}) = (c_{1i}, c_{2i})$, we first generate $\hat{V}_i = v$ from the distribution of $(V|T_1 > c_{1i}, T_2 > c_{2i})$ (whose distribution function is just the F_1 defined in Corollary 1(1)) using the inverse CDF method. We

generate $\hat{U}_i$ from the distribution of $(U|V = v, T_1 > c_{1i}, T_2 > c_{2i})$ using the result presented in Theorem 4. It is easily seen that in this way we can generate a random pair $(\hat{U}_i, \hat{V}_i)$ which follows $\{(U, V)|T_1 > c_{1i}, T_2 > c_{2i}\}$ distribution asymptotically.

(b) When (T_{1i}, T_{2i}) is singly censored and $T_{1i} > c_{1i}$, $T_{2i} = t_{2i}$, we first generate $\hat{V}_i$ from the distribution of $(V|T_1 > c_{1i}, T_2 = t_{2i})$ (whose distribution function is just the F_2 defined in Corollary 1(2)) to obtain $\hat{V}_i = v$. We estimate U_i by

$$\hat{U}_i = \frac{q_{\hat{\theta}}(v) - q_{\hat{\theta}}\{\hat{S}_2(t_{2i})\}}{q_{\hat{\theta}}(v)}.$$

(c) When (T_{1i}, T_{2i}) is singly censored and $T_{1i} = t_{1i}$, $T_{2i} > c_{2i}$, we first generate $\hat{V}_i$ from the distribution of $(V|T_1 = t_{1i}, T_2 > c_{2i})$ (whose distribution function is just the F_3 defined in Corollary 1(3)) to obtain $\hat{V}_i = v$. We estimate U_i by

$$\hat{U}_i = \frac{q_{\hat{\theta}}\{\hat{S}_1(t_{1i})\}}{q_{\hat{\theta}}(v)}.$$

(d) When (T_{1i}, T_{2i}) is uncensored in both components, i.e., $T_{1i} = t_{1i}$, $T_{2i} = t_{2i}$, we estimate U_i and V_i by

$$\hat{U}_i = \frac{q_{\hat{\theta}}(\hat{S}_1(t_{1i}))}{q_{\hat{\theta}}\{\hat{S}(t_{1i}, t_{2i})\}} \quad \text{and} \quad \hat{V}_i = \hat{S}(t_{1i}, t_{2i}),$$

respectively.

We repeat the above imputation procedures to generate m imputed data sets of size n.

3. Test step. For each complete data sample obtained from the previous data imputation step, we conduct the Z-test proposed in Section 2, and combine the test results for m ($m \geq 3$) imputed samples as described in Rubin (1987). The detailed procedure is: from m complete data sets and m values for $\hat{Z}$, $\hat{Z}_{\star 1} \ldots \hat{Z}_{\star m}$, use the test statistic $T_m = \bar{Z}_m/\sqrt{V_m}$, where $\bar{Z}_m = \sum_{l=1}^{m} \hat{Z}_{\star l}/m$ and $V_m = (1 + 1/m)B_m + 1/n$, with $B_m = \sum_{l=1}^{m}(\hat{Z}_{\star l} - \bar{Z}_m)^2/(m-1)$. The p value associated with T_m is $P_m = 2[1 - F_w(|T_m|)]$, where F_w is the distribution function of a t random variable with $w = (m-1)[V_m/(1 + m^{-1})B_m]^2$ df.

If the combined p value is less than 0.05, we reject the null hypothesis that the data can be modelled by the assumed Archimedean copula. Rubin (1987) and Li, Raghunathan and Rubin (1991) justified their way of combining test statistics by deriving the procedure theoretically. They also demonstrated their method via extensive simulation studies.

Table 1. Percentage of rejection of two different null hypotheses using statistics at the 5% level when the data is uncensored ($N = 300$ based on 1000 repetitions); the Z_n column lists our Z-test results and the S_n column lists the results of the test proposed by Genest, Quessy, and Rémillard (2006).

Family	τ	H_0 : The Clayton Model No Censoring		H_0 : The Frank Model No Censoring	
True Model		Z_n	S_n	Z_n	S_n
The Clayton model	0.3	4.1	6.6	83.6	95.0
	0.5	5.0	7.0	100.0	100.0
	0.7	5.3	2.8	100.0	100.0
The Frank model	0.3	67.1	88.0	1.9	5.4
	0.5	99.9	100.0	4.5	4.7
	0.7	100.0	100.0	5.7	4.7
The Hougaard model	0.3	94.5	99.4	11.2	45.0
	0.5	100.0	100.0	47.1	81.0
	0.7	100.0	100.0	83.0	96.0

It can be easily shown that the imputation step does produce complete imputed samples $(\hat{U}_i, \hat{V}_i)$ which follow the joint distribution of (U, V) asymptotically. A reviewer of this paper has raised a question about how to justify the previous imputation procedure, specifically, why the "pseudo observations" from MI can still have the uncorrelated property (as stated in Theorem 1) under the null hypothesis? To answer this question, we have proved the following result.

Theorem 5. *Let (T_{1i}, T_{2i}) for $i \in \{1, \ldots, n\}$ be independently identically distributed with a distribution that can be modelled by an Archimedean copula. Let (T_{1i}, T_{2i}) be subject to independent right-censoring by a censoring vector (C_{1i}, C_{2i}) that follows a bivariate continuous distribution. Under suitable regularity conditions, $(\hat{U}_i, \hat{V}_i)$ and $(\hat{U}_j, \hat{V}_j)$ defined in the above MI procedure are asymptotically independent for $i \neq j$, $i, j \in \{1, \ldots, n\}$.*

4. Simulation Studies

In this section, we demonstrate the proposed test procedures in simulation studies. The bivariate data were first generated from the Clayton model and the Frank model with unit exponential marginal distributions. Our simulation studies concern uncensored data and censored data. For uncensored bivariate data, we estimated the unknown parameter by inverting the sample estimate of τ (Genest and Rivest (1993)), i.e., $\hat{\theta} = g^{-1}(\hat{\tau})$, where g is a function such that $\tau = g(\theta)$ (for the Clayton model, $\tau = \alpha/(\alpha + 2)$; for the Frank model, $\tau = 1 + 4\{D_1(\beta) - 1\}/\beta$, where D_1 is the Debye function defined by $D_1(\beta) = \int_0^\beta t/(e^t - 1)dt/\beta$).

Table 2. Percentage of rejection of two different null hypotheses using statistics at the 5% level when the data is under 30% censoring (here censoring means that at least one component is censored, the percentages are obtained based on 1,000 repetitions).

Family	τ	H_0 : The Clayton Model		H_0 : The Frank Model	
True Model		$N = 100$	$N = 300$	$N = 100$	$N = 300$
The Clayton model	0.3	1.0	2.0	2.0	10.7
	0.5	2.1	3.0	5.6	43.5
	0.7	4.3	5.3	14.8	73.4
The Frank model	0.3	39.4	76.8	2.2	2.0
	0.5	77.8	100.0	3.0	2.2
	0.7	91.6	100.0	4.4	3.2
The Hougaard model	0.3	47.4	84.6	12.0	19.6
	0.5	87.2	99.6	32.0	67.0
	0.7	97.8	100.0	58.2	94.2

Simulation results are presented in Table 1, and suggest that our test performs quite well. Comparing it with the one proposed by Genest, Quessy, and Rémillard (2006), we find that their test for uncensored data is slightly more powerful than ours when the dependence (measured by Kendall's τ) is relatively weak; when the dependence gets stronger, both tests are powerful in detecting departure from the null model.

For censored bivariate data we used the Brown estimator (Brown et al. (1974)) to estimate Kendall's τ, following Wang and Wells (2000b). The bivariate censoring vectors were generated to follow exponential marginal distributions with mean 5. With this, about 30% of pairs will have at least one component censored (either singly or doubly censored). For our test, we chose $m = 10$ when generating imputed data; if the combined p value was less than 0.05, we rejected the null hypothesis. Results are presented in Table 2. There we can see that the proposed test procedure achieved about 5% significance level when the null model and the true model were the same. When the sample size was large ($N = 300$), our test had quite good power in detecting departures from the null model when the true model was different, except when we fit the data from the Frank model by the Clayton model. In this situation, a larger sample size (for example $N = 400$) was needed to achieve good power (according to other simulation studies we have conducted and which are not reported here). Generally speaking, the power of our test increased with sample size.

In our simulation studies, we replaced the unknown quantities (the unknown parameter and the survivor functions) with their sample counterparts. As a consequence, the resulting test may not be efficient. To get better test results, we can choose more efficient parameter estimates in the assumed Archimedean copula model (such as the estimator proposed by Shih and Louis (1995)). When the Dabrowska estimator is improper (it can be negative as pointed out by Wang and Wells (2000b)), we can estimate the joint survivor function using the assumed copula structure and the estimated marginal survivor functions. To obtain a more precise estimate of the bivariate survival function under independent right censoring, we can also try other estimators such as the Prentice-Cai estimator (Prentice and Cai (1992)). We have found that the proposed test achieves better power if the estimates of the unknown quantities in our test statistic are more efficient (accurate).

It should be emphasized that our test for uncensored data is much simpler than the one proposed by Genest, Quessy, and Rémillard (2006); we can obtain an explicit p value from the normal table based on our test statistic, while their test relies on a parametric bootstrap procedure to determine its p value. Another advantage of our test procedure is that it can be extended to a test for censored data, while theirs does not seem to have an easy extension to censored data situation.

5. An Illustrative Example: the Diabetic Retinopathy Study.

This study examined the effectiveness of laser photocoagulation for delaying the onset of blindness (visual acuity less than 5/200 at two successive visits) in patients with diabetic retinopathy. The original data set includes 197 patients, either with adult-onset diabetes, or with juvenile-onset diabetes. One eye of each patient was randomly selected for photocoagulation treatment, the other eye was untreated. Manatunga and Oakes (1999) have shown that the data set that only includes patients with adult onset diabetes can be fit well using the Clayton model based on the diagnostic plot proposed by Oakes (1989). The data set consists of 83 patients, 14 experienced failure in both eyes, 40 experienced failure in one eye, and 29 experienced no failure. Using the nonparametric estimator proposed by Brown et al. (1974), one finds $\hat{\tau} = 0.21$. We performed our second test to check the goodness-of-fit of the Clayton model and obtained a p value of 0.65, with mean correlation coefficient 0.07 for U and V based on ten complete imputed data sets; this is insignificant. We reach the same conclusion as stated in Manatunga and Oakes (1999): there is not enough evidence to reject the Clayton model assumption. Applying our test procedure to check the Frank model assumption, we obtain a p value of 0.08 with mean correlation coefficient -0.23 for U and V based on complete imputed data sets. Both tests yield insignificant p-values, and hence the data can be modelled by either model; the Clayton model seems to be a better fit.

6. Discussion

In this paper, we have proposed two goodness-of-fit tests for Archimedean copula models, one for uncensored data and the other one for right censored bivariate data. In our simulation studies, we find that both procedures work quite well. Both tests are quite simple to implement and explicit p values are provided when testing the null hypothesis.

The ideas of our new tests are the same, we use the properties of (U, V) instead of the properties of (T_1, T_2) to check the model assumption. By doing so, we avoid deriving the asymptotic distribution of test statistics such as the one proposed by Wang and Wells (2000a) or the one proposed by Genest, Quessy, and Rémillard (2006). As the asymptotic distribution of their test statistics are intractable for both uncensored and censored data; see Wang and Wells (2000a) or Genest, Quessy, and Rémillard (2006) for a more detailed discussion.

The distributional results of (U, V) given different censoring patterns are important because they provide us with insights to the structure of Archimedean copula models under right-censoring. The usefulness of our results has been shown in our data imputation step: we can impute missing bivariate data from a specified Archimedean copula model by simply applying the inverse CDF approach. With imputed data sets in hand, one can conduct "standard" complete data analyses and the test or estimation results based on imputed data sets can be combined. We expect that the proposed test procedures will be useful in correlation studies using Archimedean copula models when censoring is present.

Acknowledgements

The author would like to express gratitude to an associate editor and the referee of this paper for a thorough review, and for valuable comments that led to substantial improvements to the paper.

References

Anderson, P., Ekstraum, C. T., Klein, J., Shu, Y. and Zhang, M. (2005). A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biometrical J.* **47**, 815-824.

Brown, B. W. M., Hollander, M. and Korwar, R. M. (1974). Nonparametric tests of independence for censored data with applications to heart transplant studies. In *Reliability and Biometry: Statistical Analysis of Life Length* (Edited by R. J. Serfling), 327-354.

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

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* **16**, 1475-89.

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

Genest, C. and MacKay, R. J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *Amer. Statist.* **40**, 280-283.

Genest, C., Quessy, J-F. and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.* **30**, 337-366.

Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.* **88**, 1034-1043.

Hawkins, D. L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *Amer. Statist.* **43**, 235-237.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**, 671-678.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and and F reference distribution. *J. Amer. Statist. Assoc.* **86**, 1065-73.

Manatunga, A. K. and Oakes, D. (1999). Parametric analysis for matched pair survival data. *Lifetime Data Anal.* **5**, 371-87.

Nelsen, R. B. (1999). *An Introduction to Copulas*, Springer-Verlag, New York.

Oakes, D. (1982). A model for association in bivariate survival data. *J. Roy. Statist. Soc. Ser. B* **44**, 414-422.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.* **84**, 487-493.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, **79**, 495-512.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley, New York.

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

Shih, J. and Louis, T. A. (1995). Inference on association parameter in copula models for bivariate survival data. *Biometrics*, **26**, 183-214.

Wang, W. and Wells, M. T. (2000a). Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.* **95**, 62-72.

Wang, W. and Wells, M. T. (2000b). Estimation of Kendall's tau under censoring. *Statist. Sinica* **10**, 1199-1218.

Department of Biostatistics, Bioinformatics & Biomathematics, Georgetown University Medical Center, Washington, DC, 20007, U.S.A.

E-mail: aw94@georgetown.edu

(Received May 2008; accepted October 2008)