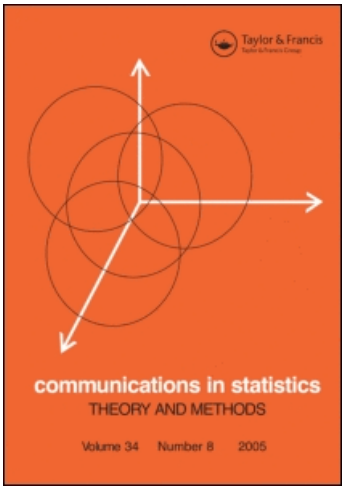

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Inference for Bivariate Survival Data by Copula Models Adjusted for the Boundary Effect

Aidong Adam Ding [a]; Weijing Wang [b]
[a] Department of Mathematics, Northeastern University, Boston, Massachusetts, USA [b] Institute of Statistics, National Chiao Tung University, Taiwan, R.O.C.

To cite this Article Ding, Aidong Adam and Wang, Weijing(2007) 'Inference for Bivariate Survival Data by Copula Models Adjusted for the Boundary Effect', Communications in Statistics - Theory and Methods, 36: 16, 2927 — 2936
To link to this Article: DOI: 10.1080/03610920701386901
URL: http://dx.doi.org/10.1080/03610920701386901

Communications in Statistics—Theory and Methods, 36: 2927–2936, 2007

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920701386901

Taylor & Francis
Taylor & Francis Group

Survival Analysis

Inference for Bivariate Survival Data by Copula Models Adjusted for the Boundary Effect

AIDONG ADAM DING[1] AND WEIJING WANG[2]

[1]Department of Mathematics, Northeastern University, Boston, Massachusetts, USA
[2]Institute of Statistics, National Chiao Tung University, Taiwan, R.O.C.

Copula models describe the dependence structure of two random variables separately from their marginal distributions and hence are particularly useful in studying the association for bivariate survival data. Semiparametric inference for bivariate survival data based on copula models has been studied for various types of data, including complete data, right-censored data, and current status data. This article discusses the boundary effect on these inference procedures, a problem that has been neglected in the previous literature. Specifically, asymptotic distribution of the association estimator on the boundary of parameter space is derived for one-dimensional copula models. The boundary properties are applied to test independence and to study the estimation efficiency. Simulation study is conducted for the bivariate right-censored data and current status data.

Keywords Copula model; Current status data; Independence test; Right-censored data; Semiparametric estimation.

Mathematics Subject Classification Primary 62N01; Secondary 62H99.

1. Introduction

Let (T_1, T_2) be a pair of correlated failure-time variables of interest. Their joint survival function can be expressed as

$$S(s, t) = \Pr(T_1 > s, T_2 > t) = C\{S_1(s), S_2(t)\}, \tag{1}$$

where $S_j(t) = \Pr(T_j > t)(j = 1, 2)$ are marginal survival functions and $C(\cdot, \cdot) : [0, 1]^2 \to [0, 1]$ is the so-called copula function which by itself is a bivariate survival (distribution) function with uniform marginals. The advantage of the

Received September 1, 2006; Accepted January 19, 2007
Address correspondence to Aidong Adam Ding, Department of Mathematics, Northeastern University, Boston, MA 02115, USA; E-mail: a.ding@neu.edu

copula representation is that the dependence structure can then be studied separately from the marginal effects. Typically, a parametric form C_θ is selected for C, with θ a real- or vector-valued parameter that is related to Kendall's tau by the formula

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

Copula models have been popular in describing the association for bivariate survival data with skewed marginal distributions. Such flexibility not only reflects in modeling but also in statistical inference. Semiparametric pseudo-likelihood estimation procedures for estimating θ have been proposed by Genest et al. (1995), Shih and Louis (1995), and Wang and Ding (2000), based on complete data, right-censored data, and current status data, respectively. In those articles, asymptotic normality of $n^{1/2}(\hat{\theta} - \theta_0)$, where $\hat{\theta}$ denotes the corresponding estimator of θ and θ_0 denotes the true parameter, has been proved when θ_0 lies in the interior of the parameter space. For some copula models, the parameter value on the boundary corresponds to important cases such as independence. However, asymptotic properties on the boundary of the parameter space have not been established yet. It should be mentioned that Genest et al. (1995) and Shih and Louis (1995) claimed that their estimators are fully efficient at independence. These arguments did not consider the boundary effect. Wang and Ding (2000) also directly applied their point estimator to test independence in analyzing a real dataset. Validity of the p-values reported in their analysis also requires further investigation.

This article is organized as follows. In Sec. 2, the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta_0)$ is derived when θ_0 is located at the boundary of the parameter space under the three different data structures mentioned above. Simulations based on right-censored data were performed to verify the theoretical derivations. In Sec. 3, we consider the problem of testing $H_0 : \theta = \theta_0$ when θ_0 is located on the boundary. Simulation results based on current status data are also presented. The efficiency property is discussed in Sec. 4. Concluding remarks are given in Sec. 5.

2. Parameter Estimation at the Boundary

Asymptotic properties of the maximum likelihood estimator on the boundary have been discussed by Moran (1971), Chant (1974), Self and Liang (1987), and Feng and McCulloch (1992). The parameter space discussed in the above articles is a subset of R^p for $p \geq 1$. The level of difficulty depends on the dimension of R^p. That is, how many of the unknown parameters are on the boundary and the correlation between the components of the maximum likelihood estimators (MLEs). In this article, we assume that (T_1, T_2) belong to the copula family in (1) and focus on a simple case in which the parameter space, denoted as Ω, is a subset of R. Denote θ^* as the lower (or upper) boundary value of Ω. That is, $\Omega = \{\theta \geq \theta^*\}$ (or $\Omega = \{\theta \leq \theta^*\}$). Examples of one-parameter copula models can be found in Joe (1993), Shih and Louis (1995), Nelsen (1999), and Genest and Rivest (2001). For many copula families, the independence case $C_\theta(u, v) = uv$ occurs at the boundary of the parameter space. Examples of such copula families are given below.

Gumbel (1960):

$$C_\theta(u, v) = e^{-[(-\log u)^\theta + (-\log v)^\theta]^{1/\theta}} \quad (\theta \geq 1);$$

Galambos (1975):

$$C_\theta(u, v) = uve^{[(-\log u)^{-\theta}+(-\log v)^{-\theta}]^{-1/\theta}} \quad (\theta \geq 0);$$

Hüsler and Reiss (1989):

$$C_\theta(u, v) = v^{\Phi(\theta^{-1}+0.5\theta \log[(-\log v)/(-\log u)])} u^{\Phi(\theta^{-1}+0.5\theta \log[(-\log u)/(-\log v)])} \quad (\theta \geq 0),$$

where $\Phi(\cdot)$ denotes the standard normal distribution function;

Joe (1993):

$$C_\theta(u, v) = 1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta} \quad (\theta \geq 1).$$

When the marginal distributions are not specified, θ can be estimated using the pseudo-likelihood approach. Specifically the semiparametric estimator $\hat{\theta}$ maximizes the pseudo log-likelihood function, $\log L(\theta, \widehat{S}_1, \widehat{S}_2)$, where $\widehat{S}_j(t)$ is a nonparametric estimator of $S_j(t) = \Pr(T_j > t)(j = 1, 2)$. Note that $\widehat{S}_j(t)$ can be the empirical estimator, the Kaplan-Meier estimator and the nonparametric MLE for complete data, right-censored data, and current status data, respectively.

Theorem 2.1. *Assume $\theta_0 = \theta^*$. When θ^* is located at the lower boundary of the parameter space, as $n \to \infty$ $n^{1/2}(\hat{\theta} - \theta^*)$ converges in distribution to the random variable $X_+ = \max(0, X) = I(X > 0)X$, where $X \sim N(0, \sigma_X^2)$. When θ^* is located at the upper boundary of the parameter space, $n^{1/2}(\hat{\theta} - \theta^*)$ converges in distribution to the random variable $X_- = \min(0, X)$.*

The normal random variable X actually has the asymptotic normal distribution of $n^{1/2}(\hat{\theta} - \theta)$ if $\theta_0 = \theta$ lies in the interior of Ω. The variance of X, denoted as σ_X^2, has been derived under three different data structures. Specifically, for complete data, $\sigma_X^2 = v^2$ given in Eq. (3) of Genest et al. (1995, p. 545); for right-censored data, $\sigma_X^2 = \tau^2$ in the last line of p. 1388 in Shih and Louis (1995); and for current status data, $\sigma_X^2 = \sigma^2$ on p. 884 of Wang and Ding (2000). We give the proof of Theorem 2.1 in the Appendix for current status data (the proof for the other two data types are very similar and thus omitted).

The finite sample distribution of $\hat{\theta}$ was examined via simulations based on right-censored data. The true survival times T_1, T_2 were generated from the Gumbel family and then the times were subjected to right censoring by two independent uniformly distributed random censoring variables, C_1 and C_2. The sample size is 400 and the censoring proportion is about 50%. The method proposed by Shih and Louis (1995) was applied to obtain $\hat{\theta}$. Based on 1,000 simulation runs, there were 509 times with $\hat{\theta} = 1$. This agrees with the statement in Theorem 2.1 that $\hat{\theta}$ follows the mixture distribution with 50% probability at the true parameter value 1. For the other situation with $\hat{\theta} \neq 1$, Theorem 2.1 states that asymptotically the estimates follow a normal distribution truncated at the center. To verify whether this is true, we did a normal probability plot. Based on those estimates $\hat{\theta} \neq 1$, we subtracted 1 from each estimate to get $x_1, x_2, \ldots, x_m$. Then we made a normal probability plot based on $x_1, x_2, \ldots, x_m$ combined with $-x_1, -x_2, \ldots, -x_m$. This plot should be close to a straight line if $\hat{\theta}$ is distributed as claimed in Theorem 2.1. We see that the normal probability plot in Fig. 1 is indeed very close to a straight

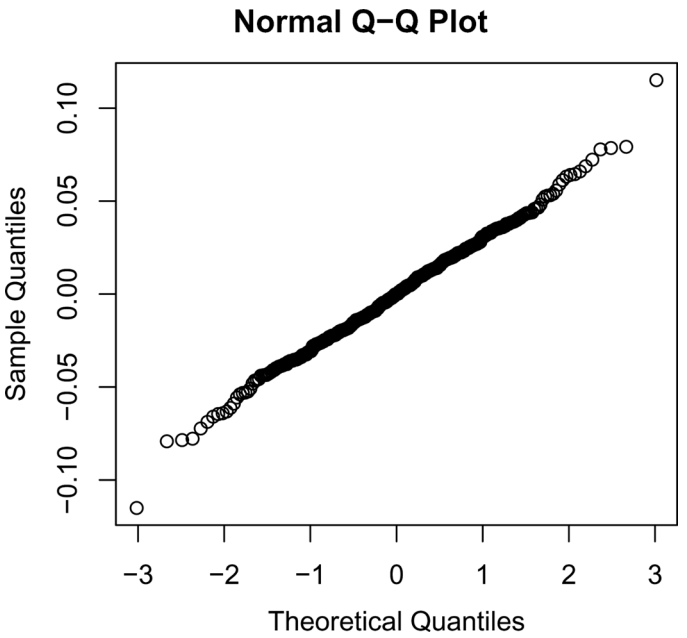

Figure 1. The normal probability plot for $\hat{\theta} - \theta$ and $\theta - \hat{\theta}$. This should be close to a straight line if the estimator is distributed as what the Theorem 2.1 stated.

line. This shows that the asymptotical result is a very good approximation for the simulation setting. We also conducted simulation with censoring rates of 20% and 80% which yield the proportions of $\hat{\theta} = 1$ equal to 55.7% and 49.7%, respectively. The normal probability plots for these two cases are similar.

3. Hypothesis Testing at the Boundary with an Application to Bivariate Current Status Data

Now we consider the problem of testing $H_0 : \theta = \theta_0$ when $\theta_0 = \theta^*$. Without loss of generality, let $H_a : \theta > \theta_0$. For the copula families where the independence corresponds to the boundary parameter value, the null hypothesis states that the two failure times are independent. Therefore, the test discussed here becomes a test of independence under the assumed copula model. Let α be the level of significance. When $\hat{\theta} = \theta_0$, obviously H_0 is accepted. Consider the decision rule such that H_0 is rejected if

$$\frac{n^{1/2}(\hat{\theta} - \theta_0)}{\hat{\sigma}_X} > z_\alpha, \tag{2}$$

where $z_\alpha > 0$ is the cut-off point satisfying $1 - \Phi(z_\alpha) = \alpha$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\hat{\sigma}_X^2$ is a consistent estimator of σ_X^2 which will be discussed later. It is easy to see that when $\alpha < 0.5$, under H_0 $\Pr\big(n^{1/2}\{\hat{\theta} - \theta_0\}/\hat{\sigma}_X > z_\alpha\big)$ converges to $\Pr(X/\sigma_X > z_\alpha) = \alpha$. Equivalently, one can construct the test by comparing the values of p-value and α. When

$\hat{\theta} - \theta_0 > 0$, the p-value can be written as

$$\Pr\big(X_+ > n^{1/2}\{\hat{\theta} - \theta_0\}\big) = \Pr\left(Z > \frac{n^{1/2}\{\hat{\theta} - \theta_0\}}{\hat{\sigma}_X}\right) = 1 - \Phi\left(\frac{n^{1/2}\{\hat{\theta} - \theta_0\}}{\hat{\sigma}_X}\right),$$

where Z is the standard normal distribution. It is easy to see that as $n \to \infty$,

$$\Pr\left(1 - \Phi\left\{\frac{n^{1/2}(\hat{\theta} - \theta_0)}{\hat{\sigma}_X}\right\} \le \alpha\right) = \alpha,$$

when $\alpha < 0.5$. The above analysis implies that when $\alpha < 0.5$, one can use the same one-sided test derived without considering the boundary effect.

Now we discuss variance estimation. There are two reasonable analytic candidates, namely $\hat{\sigma}^2_X(\hat{\theta})$ and $\hat{\sigma}^2_X(\theta^*)$. The two estimators only differ in the plugged-in estimator of θ. Using $\hat{\sigma}^2_X(\hat{\theta})$ is like inverting the confidence interval directly for hypothesis testing without adjusting for the boundary effect. Under H_0, both variance estimators are consistent but the convergence of $\hat{\sigma}^2_X(\hat{\theta})$ to the true variance σ^2_X is slower than $\hat{\sigma}^2_X(\theta_0)$. Furthermore, when we reject H_0, the p-value corresponds to a small value and, equivalently, $\hat{\theta} - \theta_0$ is large which are the cases where $\hat{\theta}$ is not a very accurate estimate of θ. Therefore the Type 1 error based on the test statistic using $\hat{\sigma}^2_X(\hat{\theta})$ is not as accurate as that using $\hat{\sigma}^2_X(\theta_0)$.

We now apply the above result to test the independence for bivariate current status data. For the copula families that independence occurs at the boundary of parameter space, testing $H_0 : \theta = \theta^*$ is equivalent to testing independence under the assumed copula model. For bivariate current status data, the formula of $\hat{\sigma}^2_X$ has been derived by Wang and Ding (2000). Alternatively, Ding and Wang (2004) proposed a nonparametric test for testing independence based on bivariate current status data. The nonparametric test is more robust under model mis-specification. However, the semi-parametric testing procedure is likely to produce a more powerful test if the underling model is correctly specified. Simulations were performed to examine our conjecture. The true failure times were generated from the Gumbel family and then current status type of data were constructed with the censoring distribution independent of (T_1, T_2). As expected, the test based on $\hat{\sigma}^2_X(\theta_0)$ produces better power than that based on $\hat{\sigma}^2_X(\hat{\theta})$. However even the size of the former test is too conservative and converges to the nominal level only at very large sample size with $n \ge 1,000$ (data not shown here). Hence, both variance estimators are not quite satisfactory.

To improve variance estimation, we propose to use bootstrapping similar to the idea used in Ding and Wang (2004). For current status data, the true failure times (T_{1k}, T_{2k}) are censored by C_k such that one only observes $\{(C_k, \delta_{1k}, \delta_{2k})\ k = 1, \ldots, n\}$, where $\delta_{jk} = T_{jk} \wedge C_k\ (j = 1, 2)$ for $k = 1, \ldots, n$. From the original data, one can generate a pseudo dataset, $\big\{\big(C_k, \delta^*_{1k}, \delta^*_{2k}\big)\, k = 1, \ldots, n\big\}$, where δ^*_{jk} is a Bernoulli random variable with probability $\widehat{F}_j(C_k)\ (j = 1, 2)$. The procedure is repeated m times. The semi-parametric estimator was computed for each pseudo dataset and let $\hat{\theta}^{(r)}$ be the estimator based on the rth bootstrapped sample. Then the proposed bootstrap variance estimator is calculated by

$$\hat{\sigma}^2_{X(b)} = \sum_{r=1}^{m} \big(\hat{\theta}^{(r)} - \theta^*\big)^2 \Big/ \sum_{r=1}^{m} I\big\{\hat{\theta}^{(r)} \ne \theta^*\big\}$$

where $I(\cdot)$ is the indicator function.

Table 1

Empirical powers of the test (2) with $\hat{\sigma}^2_{X(b)}$ and the test of Ding and Wang (2004) based on 4,000 replications. The first column lists the sample size (n), the second column lists the prevalence rate (P.L.) of each simulated data sets. The data were generated for the Gumbel family with different levels of Kendall's τ listed on the top row. In each cell, the first number is the empirical probability of rejecting H_0 using test (2) with $\hat{\sigma}^2_{X(b)}$ and the number in the parenthesis is that using the test of Ding and Wang (2004)

		Correlation τ							
n	*P.L.*	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35
200	≈20%	0.040	0.148	0.434	0.656	0.853	0.949	0.983	0.995
		(0.041)	(0.130)	(0.421)	(0.652)	(0.838)	(0.943)	(0.981)	(0.994)
	≈50%	0.042	0.156	0.343	0.578	0.763	0.914	0.976	0.996
		(0.049)	(0.172)	(0.355)	(0.605)	(0.784)	(0.923)	(0.981)	(0.997)
400	≈20%	0.045	0.279	0.662	0.912	0.986	0.999	1.000	1.000
		(0.039)	(0.237)	(0.596)	(0.883)	(0.982)	(0.999)	(1.000)	(1.000)
	≈50%	0.044	0.222	0.554	0.838	0.966	0.997	1.000	1.000
		(0.048)	(0.211)	(0.533)	(0.817)	(0.963)	(0.997)	(1.000)	(1.000)

We conducted simulations to compare test (2) using $\hat{\sigma}^2_{X(b)}$ with the test with highest power in Ding and Wang (2004) at nominal level $\alpha = 0.05$. The simulation setup is the same as that mentioned earlier. We evaluated cases with two different sample sizes: $n = 200$ and $n = 400$; two different prevalence rates $\Pr(\delta_j = 1)$: 20% and 50%; and different levels of Kendall's $\tau : 0, 0.05, 0.10, \ldots, 0.35$. For each combination, 4,000 simulation runs were conducted and the results were summarized in Table 1.

From Table 1, we can see that the test (2) using $\hat{\sigma}^2_{X(b)}$ has correct size (i.e., Type I error rate under the null hypothesis of $\tau = 0.00$). And the size of the test (2) becomes closer to the nominal level $\alpha = 0.05$ when the sample size increases from $n = 200$ to $n = 400$. The power of the test (2) using $\hat{\sigma}^2_{X(b)}$ is higher than the power of Ding and Wang (2004)'s test in most cases. When $n = 200$ and prevalent level is 50%, the test (2) using $\hat{\sigma}^2_{X(b)}$ do have lower power. That seems due to its lower size than Ding and Wang (2004)'s tests. As sample size increases, its size increases to the nominal level and its power also increases.

4. Efficiency of the Estimator

Genest et al. (1995) and Shih and Louis (1995) claimed that their semi-parametric estimators are fully efficient under independence (i.e., $S(t_1, t_2) = S_1(t_1)S_2(t_2)$). They showed that the semi-parametric estimator $\hat{\theta}$ and the efficient estimator θ both have asymptotical normal distributions with the same variance under independence. Here, $\tilde{\theta}$ is the maximum likelihood estimator obtained when S_1 and S_2 are completely specified. Note that $\tilde{\theta}$ is an efficient estimator since its variance achieves the Cramer-Rao lower bound asymptotically (except at the boundary value which is a set of Lebesgue measure zero). The analysis in Genest et al. (1995) and Shih and Louis (1995), however, did not consider the boundary situation. Here we give our

explanations. When the independence case occurs at the boundary of the parameter space, neither $\hat{\theta}$ nor $\tilde{\theta}$ are asymptotic normally distributed. However, if $\theta_0 = \theta^*$ under independence, then $n^{1/2}(\tilde{\theta} - \theta_0)$ still has the same asymptotic distribution as $n^{1/2}(\hat{\theta} - \theta_0)$, both of which converge to the mixture distribution X_+ as $n \to \infty$. Therefore, the efficiency claims are still true if we define efficiency as having the same asymptotic distribution as $\tilde{\theta}$ which is the best we can hope for in practice.

Now we show similar efficiency property for the estimator proposed by Wang and Ding (2000) based on bivariate current status data. On p. 891 of Wang and Ding (2000), we have the expression:

$$Q(\theta_0, S_1, S_2, c, \delta_1, \delta_2) = \frac{\partial}{\partial\theta} l(\theta, S_1(c), S_2(c), \delta_1, \delta_2)|_{\theta=\theta_0} - \tilde{l}(c, \delta_1, S_1, G, \psi_1) - \tilde{l}(c, \delta_2, S_2, G, \psi_2).$$

The last two terms appear since the survival distribution functions $S_j(\cdot)\,(j = 1, 2)$ are estimated by the corresponding NPMLEs. Conditional on the censoring time C_i and $\delta_{2,i}\,(i = 1, \ldots, n)$, it is easy to show that the covariance of the first term with $\tilde{l}(c, \delta_1, S_1, G, \psi_1)$ is zero. Hence, the unconditional covariance is also zero. Similar arguments apply to the covariance of the first term with the last term. Therefore,

$$\mathrm{Cov}\left\{ \frac{\partial}{\partial\theta} l(\theta, S_1(c), S_2(c), \delta_1, \delta_2)|_{\theta=\theta_0}, \tilde{l}(c, \delta_j, S_j, G, \psi_j) \right\} = 0 \quad (j = 1, 2).$$

Consequently, the variance of $Q(\theta_0, S_1, S_2, c, \delta_1, \delta_2)$ is bounded below by

$$\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}\left(\frac{\partial}{\partial\theta} l\{\theta, S_1(c), S_2(c), \delta_1, \delta_2)|_{\theta=\theta_0}\} \right).$$

Direct calculations show that $\tilde{l}(c, \delta_j, S_j, G, \psi_j)\} = 0\,(j = 1, 2)$ under independence no matter whether θ_0 lies in the interior or on the boundary. Thus, replacing marginals S_1 and S_2 by their univariate NPMLEs does not add to asymptotic variance only at the independence case. Hence, Wang and Ding's estimator also has the same distribution as $\tilde{\theta}$ (the MLE assuming marginals S_1 and S_2 are known) under independence. Therefore the semi-parametric estimator is also efficient under independence for current status data similar to the cases for complete data and for right-censored data.

5. Concluding Remarks

In this article, we discuss the boundary problem for one-parameter copula models with $\theta \in R$. Examples of two-parameter copula families can be found in Genest and Rivest (1993, 2001) and Nelsen (1999). The pseudo-likelihood estimation procedure can be easily extended to multiparameter cases as illustrated in Sec. 4 of Genest et al. (1995). For the boundary problem on higher dimensions, one can apply the results of Self and Liang (1987) based on maximum likelihood estimation. Although the techniques can be directly applied under the context of pseudo-likelihood estimation, derivations of the high-dimensional mixture distribution require a lot of analytical work. Consider the example of $\theta = (\theta_1, \theta_2)^T \in \Omega = \Omega_1 \times \Omega_2 \subset R^2$. The case 2 of Self and Liang (1987) discusses the situation when the true value of

θ_1 is located on the boundary of Ω_1 but that of θ_2 lies in the interior of Ω_2. Their case 3 discusses the situation when both of $\theta_j\,(j=1,2)$ are located on the boundary of Ω_j. In the special case such that the two parameters are orthogonal $\left(\text{that is, } \frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2}\log L(\theta,\widehat{S}_1,\widehat{S}_2)=0\right)$, it is easy to derive the asymptotic distribution of $n^{1/2}(\hat{\theta}-\theta_0)$ from marginal analysis. However, in general, explicit derivations are complicated.

One should be cautious when applying the iterative algorithms as illustrated in Genest et al. (1995), Shih and Louis (1995), and Wang and Ding (2000). If the maximum occurs on the boundary, the iterative algorithms will break down. For the one-parameter case this can be resolved by adding a step to check the sign of $U\left(\theta^*,\widehat{S}_1,\widehat{S}_2,H_n\right)=\frac{1}{n}\frac{\partial}{\partial\theta}\log L\left(\theta^*,\widehat{S}_1,\widehat{S}_2\right)$. Specifically, for lower boundary value θ^*, if $U\left(\theta^*,\widehat{S}_1,\widehat{S}_2,H_n\right)<0$ then set $\hat{\theta}=\theta^*$; otherwise the algorithm will find a solution in the interior of the parameter space. Implementation of the maximization algorithm for the multiple parameters case, however, is much harder although it is theoretically feasible.

If the independence case happens at the boundary of a multi-dimensional parameter space, it becomes difficult to test independence under the semiparametric framework due to the aforementioned difficulties involving the algorithm and asymptotic distribution. Nonparametric tests for independence, which do not make any model assumption, have been proposed by Hsu and Prentice (1996) and Shih and Louis (1996) for right-censored data and Ding and Wang (2004) for current status data. However, when we know the copula family, these tests usually have lower asymptotic power than the semiparametric tests for independence as shown by the simulation study.

Appendix

Asymptotic Distribution of $n^{1/2}(\hat{\theta}-\theta_0)$ When $\theta_0=\theta^*$

We follow the notation of Wang and Ding (2000). Under the copula model in (1), Wang and Ding proposed a semiparametric method for estimating θ based on current status data of the form, $\{(C_i,\delta_{1i},\delta_{2i})\,(i=1,\ldots,n)\}$, which are *iid* replications of $\{C,\delta_1=I(T_1\le C),\delta_2=I(T_2\le C)\}$, where C is the common monitoring time for the pair. It is assumed that C is independent of (T_1,T_2). The proposed estimator $\hat{\theta}$ maximizes the pseudo log-likelihood function, $\log L\left(\theta,\widehat{S}_1,\widehat{S}_2\right)$, where $\widehat{S}_j(t)$ is the nonparametric MLE of $S_j(t)=\Pr(T_j>t)\,(j=1,2)$ discussed in Groeneboom and Wellner (1992, pp. 66–67). The corresponding score equation is defined as

$$U\left(\theta,\widehat{S}_1,\widehat{S}_2,H_n\right)=\frac{1}{n}\frac{\partial}{\partial\theta}\log L\left(\theta,\widehat{S}_1,\widehat{S}_2\right)=0,$$

where H_n is the empirical distribution of the observed data, and the derivative is taken from the appropriate side if θ is close to the boundary value of Ω, denoted as θ^*. If the maximum occurs in the interior region of Ω, $\hat{\theta}$ is the solution to the score equation. However, the equation does not have a solution if the maximum occurs on the boundary. In such a case, $\hat{\theta}=\theta^*$, $U\left(\theta^*,\widehat{S}_1,\widehat{S}_2,H_n\right)<0$ and the iterative algorithm cannot be applied.

Applying the results in Wang and Ding (2000, p. 891), one can show that $n^{1/2}U\big(\theta_0, \widehat{S}_1, \widehat{S}_2, H_n\big)$ is asymptotically normal with mean zero and variance $Q(\theta_0, S_1, S_2, C, \delta_1, \delta_2)$ defined in (7) (Wang and Ding, 2000, p. 884, 891). Therefore, as $n \to \infty$,

$$\Pr\big(n^{1/2}U\big(\theta_0, \widehat{S}_1, \widehat{S}_2, H_n\big) > 0\big) = \Pr\big(n^{1/2}U\big(\theta_0, \widehat{S}_1, \widehat{S}_2, H_n\big) < 0\big) = 0.5.$$

When $n^{1/2}U\big(\theta_0, \widehat{S}_1, \widehat{S}_2\big) < 0$, the pseudo log-likelihood is maximized at $\theta_0 = \theta^*$, it follows that $n^{1/2}(\hat{\theta} - \theta_0) = 0$. When $n^{1/2}U\big(\theta_0, \widehat{S}_1, \widehat{S}_2\big) > 0$, the score equation has a solution. A Taylor expansion can be conducted at the solution $\hat{\theta}$, and the analysis in Wang and Ding (2000) implies that

$$n^{1/2}(\hat{\theta} - \theta_0) =^a -[V(\theta_0, S_1, S_2, H)]^{-1} n^{1/2} U\big(\theta_0, \widehat{S}_1, \widehat{S}_2, H_n\big),$$

where $V(\theta_0, S_1, S_2, H)$ is defined in Sec. 2.3 (Wang and Ding, 2000). Therefore the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta_0)$ is a mixture distribution with probability 0.5 at the mass point, 0, and probability 0.5, being the positive half of the mean-zero normal random variable X with variance

$$\sigma^2 = [V(\theta_0, S_1, S_2, H)]^{-2} \operatorname{Var}\{Q(\theta_0, S_1, S_2, C, \delta_1, \delta_2)\}. \tag{A.1}$$

The same techniques can be applied under the two other data structures. Specifically, the score equations derived for complete data (Genest et al., 1995) and for right-censored data (Shih and Louis, 1995) are also mean-zero normal variables. Similar arguments can be applied to show that the corresponding asymptotic distributions at the boundary are also mixture distributions.

References

Chant, D. (1974). On asymptotic tests of composite hypothesis in nonstandard conditions. *Biometrika* 61:291–298.

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* 65:141–151.

Ding, A. A., Wang, W. (2004). Testing independence for bivariate current status data. *J. Amer. Statist. Assoc.* 99:145–155.

Feng, Z. F., McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statist. Probab. Lett.* 13:325–332.

Galambos, J. (1975). Order statistics of samples from multivariate distributions. *J. Amer. Statist. Assoc.* 70:674–680.

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

Genest, C., Rivest, L.-P. (2001). On the multivariate probability integral transformation. *Statist. Probab. Lett.* 53:391–399.

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

Groeneboom, P., Wellner, J. A. (1992). *Information Bounds and Non-Parametric Maximum Likelihood Estimation.* Boston: Birkhäuser.

Gumbel, E. J. (1960). Distributions des valeurs extremes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris* 9:171–173.

Hsu, L., Prentice, R. L. (1996). A generalization of the Mantel-Haenszel tests to bivariate failure time data. *Biometrika* 83:905–911.

Hüsler, J., Reiss, R.-D. (1989). Maxima of normal random vectors between independence and complete dependence. *Statist. Probab. Lett.* 7:283–286.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *J. Multivariate Anal.* 46:267–282.

Moran, P. A. P. (1971). Maximum likelihood estimation in non-standard conditions. *Proc. Cambridge Philos. Soc.* 70:441–450.

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

Self, S. G., Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Amer. Statist.* 82:605–610.

Shih, J. H., Louis, T. A. (1995). Inference on the association parameter in copula models for bivariate survival data. *Biometrics* 51:1384–1399.

Shih, J. H., Louis, T. A. (1996). Tests of independence for bivariate survival data. *Biometrics* 52:1440–1449.

Wang, W., Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika* 87:879–893.