Estimation and Goodness of Fit for Multivariate Survival Models Based on Copulas

by

Yildiz Elif Yilmaz

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We provide ways to test the fit of a parametric copula family for bivariate censored data with or without covariates. The proposed copula family is tested by embedding it in an expanded parametric family of copulas. When parameters in the proposed and the expanded copula models are estimated by maximum likelihood, a likelihood ratio test can be used. However, when they are estimated by two-stage pseudolikelihood estimation, the corresponding test is a pseudolikelihood ratio test. The two-stage procedures offer less computation, which is especially attractive when the marginal lifetime distributions are specified nonparametrically or semiparametrically. It is shown that the likelihood ratio test is consistent even when the expanded model is misspecified. Power comparisons of the likelihood ratio and the pseudolikelihood ratio tests with some other goodness-of-fit tests are performed both when the expanded family is correct and when it is misspecified. They indicate that model expansion provides a convenient, powerful and robust approach.

We introduce a semiparametric maximum likelihood estimation method in which the copula parameter is estimated without assumptions on the marginal distributions. This method and the two-stage semiparametric estimation method suggested by Shih and Louis (1995) are generalized to regression models with Cox proportional hazards margins. The two-stage semiparametric estimator of the copula parameter is found to be about as good as the semiparametric maximum likelihood estimator. Semiparametric likelihood ratio and pseudolikelihood ratio tests are considered to provide goodness of fit tests for a copula model without making parametric assumptions for the marginal distributions. Both when the expanded family is correct and when it is misspecified, the semiparametric pseudolikelihood ratio test is almost as powerful as the parametric likelihood ratio and pseudolikelihood ratio tests while achieving robustness to the form of the marginal distributions. The methods are illustrated on applications in medicine and insurance.

Sequentially observed survival times are of interest in many studies but there are difficulties in modeling and analyzing such data. First, when the duration of followup is limited and the times for a given individual are not independent, the problem of induced dependent censoring arises for the second and subsequent survival times. Non-identifiability of the marginal survival distributions for second and later times is another issue, since they are observable only if preceding survival times for an individual are uncensored. In addition, in some studies, a significant proportion of individuals may never have the first event. Fully parametric models can deal with these features, but lack of robustness is a concern, and methods of assessing fit are lacking. We introduce an approach to address these issues. We model the joint distribution of the successive survival times by using copula functions, and provide semiparametric estimation procedures in which copula parameters are estimated without parametric assumptions on the marginal distributions. The performance of semiparametric estimation methods is compared with some other estimation methods

in simulation studies and shown to be good. The methodology is applied to a motivating example involving relapse and survival following colon cancer treatment.

Acknowledgements

I want to express my deepest gratitude to my supervisor Prof. Jerald Lawless for his invaluable support, guidance, encouragement and insight throughout this research work. This thesis would not have been possible without his guidance, advices, criticisms and most importantly without his care.

I would like to thank my thesis committee members Prof. Christian Genest, Prof. David Matthews, Prof. Grace Yi and Prof. John Flanagan for their insightful comments.

I wish to thank Candemir Cigsar for his support during my study at university and my parents and grandparents for a lifetime of encouragement.

To the memory of my grandfather, Burhan Okan.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multivariate Lifetime Distributions and Copula Models

Multivariate lifetime data analysis is necessary in various settings (Lawless, 2003; Hougaard, 2000). Multivariate lifetime data includes parallel clustered data in which each subject has more than one failure time which are observed in parallel or simultaneously and do not satisfy any order restrictions; for example, times to occurrence of a disease in paired organs within individuals as in the Diabetic Retinopathy Study data given in Section 1.1.1 or times to disease onset or death in related individuals (Hougaard et al., 1992; Hsu and Gorfine, 2006). It also includes sequential data in which sequences of survival times, observed one after the other. An example is the times between successive events for an individual such as entry to the stages of a two-stage disease process as in the Colon Cancer data given in Section 1.1.2.

The objectives of this thesis are to develop some new methods for use with bivariate lifetime data. In particular, we will provide methods for tests of fit, for semiparametric estimation, and for the analysis of sequentially observed data. These topics are discussed in Section 1.6, but first we provide a review of models and methods for bivariate lifetimes.

We focus on the case of bivariate lifetimes but the approaches developed in the thesis can also be applied to settings with three or more lifetimes. In this chapter, approaches to model bivariate lifetime data are explained, specifically the general description of copula models. Some important models are summarized, and a review of statistical methods for copula models and previous work on copula model selection and goodness of fit tests is given.

1.1.1 Diabetic Retinopathy Study Data

The Diabetic Retinopathy Study (DRS) was begun in 1971 to study the effectiveness of laser photocoagulation treatment in delaying the onset of blindness in patients with diabetic retinopathy. Diabetic retinopathy occurs in diabetic persons and causes blindness. Huster et al. (1989) gives some important details of the study. The patients were eligible for the study if they had diabetic retinopathy and visual acuity of 20/100 or better in both eyes. One eye of each patient was randomly selected for treatment and the other eye was observed without treatment. The variable used to assess the treatment effect was the time to occurrence of visual acuity less than 5/200 at two consecutively completed 4-month follow-ups. A 50% sample of the high-risk patients as defined by DRS criteria gives a subset of $n = 197$ subjects. It is important to understand whether there is an effect of the laser photocoagulation treatment. Many authors such as Huster et al. (1989), Glidden and Self (1999), He and Lawless (2003, 2005) and Romeo et al. (2006) analyzed this data set.

1.1.2 Colon Cancer Data

A clinical trial was conducted to assess the effectiveness of a therapy with levamisole plus fluorouracil compared to a placebo with respect to colon cancer patients' cancer recurrence and survival. Moertel et al. (1990) and Lin et al. (1999) give some information about the study. Colon cancer is a common cancer type. When the diagnosis is made at a sufficiently early stage, all apparent diseased tissue can be surgically removed. The patients who have regional nodal involvement that is clinically completely resected are referred to as having Duke's Stage C disease. Some patients have residual cancer existing in an occult and probably in microscopic stage, which leads to recurrence of disease and death within 5 years. In this randomized clinical trial on Duke's Stage C patients there were 315 patients assigned to the placebo group and 304 patients assigned to the levamisole plus fluorouracil therapy group. Maximum follow-up time was approximately 9 years. It is important to assess whether there are effects of the therapy with levamisole plus fluorouracil on the time from study registration to cancer recurrence, and if the cancer recurs, on the time from recurrence to death. Some authors such as Moertel et al. (1990), Lin et al. (1999) and He and Lawless (2003) analyzed this data set.

1.2 Bivariate Lifetime Distributions

Let T_1, T_2 be lifetime variables of an individual which may not be independent. The bivariate distribution and survivor functions for $t_1 \geq 0$ and $t_2 \geq 0$ are defined as

$$F(t_1, t_2) = Pr(T_1 \leq t_1, T_2 \leq t_2) \tag{1.1}$$

and

$$S(t_1, t_2) = Pr(T_1 \geq t_1, T_2 \geq t_2), \tag{1.2}$$

respectively. For continuous lifetime variables T_1, T_2, the bivariate survivor function can be expressed in terms of the distribution functions as follows:

$$S(t_1, t_2) = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2). \tag{1.3}$$

The marginal distribution fuctions of T_1 and T_2 are $F_1(t_1) = F(t_1, \infty)$ and $F_2(t_2) = F(\infty, t_2)$ and the marginal survivor functions are $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$, respectively. The hazard rate of the conditional distribution of T_i given $T_j = t_j$ is

$$\lambda_{ij}(t_i|t_j) = \lim_{\Delta t \to 0} \frac{Pr(T_i < t_i + \Delta t | T_i \geq t_i, T_j = t_j)}{\Delta t} = \frac{\partial^2 S(t_i, t_j)/\partial t_i \partial t_j}{-\partial S(t_i, t_j)/\partial t_j} \tag{1.4}$$

for $t_i > t_j$, $i \neq j$ and $i, j = 1, 2$. The hazard rate of the conditional distribution of T_i given $T_j \geq t_j$ is

$$\lambda'_{ij}(t_i|t_j) = \lim_{\Delta t \to 0} \frac{Pr(T_i < t_i + \Delta t | T_i \geq t_i, T_j \geq t_j)}{\Delta t} = \frac{-\partial S(t_i, t_j)/\partial t_i}{S(t_i, t_j)} \tag{1.5}$$

for $i \neq j$ and $i, j = 1, 2$.

Suppose the parallel clustered lifetimes (T_{1i}, T_{2i}) of a random sample of individuals $i = 1, ..., n$ have common continuous survivor function $S(t_1, t_2)$ and potential right censoring times (C_{1i}, C_{2i}) assumed to be independent of the lifetimes. Let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_{1i}), \min(T_{2i}, C_{2i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$ be the observed data and its censoring indicators, respectively. The likelihood function is (Lawless, 2003, Section 11.2)

$$L = \prod_{i=1}^{n} f(t_{1i}, t_{2i})^{\delta_{1i}\delta_{2i}} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right]^{(1-\delta_{1i})\delta_{2i}} S(t_{1i}, t_{2i})^{(1-\delta_{1i})(1-\delta_{2i})} \tag{1.6}$$

where $f(t_1, t_2) = \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2}$.

When the sequence of lifetimes (T_{1i}, T_{2i}), observed in order, represents the times between a sequence of events and C_i denotes the censoring time (total followup time) for individual i, $i = 1, ..., n$, there may be three different types of observations: (i) T_{1i} is not observed, i.e. $t_{1i} = C_i$; (ii) $T_{1i} = t_{1i}$ is observed but T_{2i} is not observed, i.e. $t_{2i} = C_i - t_{1i}$; (iii) both $T_{1i} = t_{1i}$ and $T_{2i} = t_{2i}$ are observed. Let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_i), \min(T_{2i}, C_i - t_{1i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$ be the observed lifetimes and their censoring indicators, respectively. Then, the likelihood function is (Lawless, 2003, Section 11.3)

$$L = \prod_{i=1}^{n} f(t_{1i}, t_{2i})^{\delta_{1i}\delta_{2i}} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} S_1(t_{1i})^{1-\delta_{1i}}. \tag{1.7}$$

Note that the censoring time for T_{1i} is C_i but since T_{2i} cannot be observed until T_{1i} has been observed, the censoring time for T_{2i} is $C_i - t_{1i}$ if $T_{1i} = t_{1i}$. If T_{1i} and T_{2i} are not independent, the censoring time for T_{2i} is not independent of T_{2i}.

When there are fixed covariates x present we denote lifetime distributions as $F(t_1, t_2|x)$, $S(t_1, t_2|x)$, $F_j(t_j|x)$, and so on. The likelihood functions (1.6) and (1.7) still apply when there are covariates, with $S_1(t_{1i})$, $f(t_{1i}, t_{2i})$ and $S(t_{1i}, t_{2i})$ replaced with $S_1(t_{1i}|x_i)$, $f(t_{1i}, t_{2i}|x_i)$ and $S(t_{1i}, t_{2i}|x_i)$, respectively.

There are two main approaches to model parallel clustered lifetime data: marginal approach and random effect models. In the marginal approach the joint distribution of T_1 and T_2 is modelled directly, as with a bivariate exponential distribution, for example. In this case the marginal distributions are usually modelled separately from the dependency structure. Copula models (Joe, 1997; Nelsen, 2006) and log-location-scale models (Lawless, 2003) are common models for the marginal approach. However, a bivariate random effect model assumes conditional independence of T_1 and T_2, given an unobserved random variable. It is often called a frailty model (Hougaard, 2000) if the conditional distributions of T_1 and T_2 have conditional proportional hazards models. Integration with respect to the unobserved variable gives the joint distribution for T_1 and T_2, but it is usually the case that certain parameters affect both the marginal distributions and association.

Methods for sequences of lifetimes are discussed in Cook and Lawless (2007, Chapter 4), in connection with the gap times between recurrent events. There are three main approaches to model sequential lifetime data: marginal models, random effect models and conditional models. Marginal models and random effect models are described in detail in the following sections. In the conditional approach, $F_1(t_1|x)$ and the conditional distribution function for T_2 given T_1, denoted $F_{2|1}(t_2|t_1, x)$ are modeled for example by using proportional hazards model or accelerated failure time models. To model the conditional distribution function $F_{2|1}$, t_1 is effectively included among the covariates for T_2. This approach is better when there are time-varying covariates. However, the marginal distribution $F_2(t_2|x)$ generally turns out to be in a complicated form and it is difficult to interpret the marginal effect of fixed covariates x on T_2. We focus in this thesis on fixed covariates. Random effect marginal models also do not always provide marginal distributions in simple form. However, for copula models or bivariate accelerated failure time models, discussed below, the marginal distributions have easily interpretable forms because these approaches allow us to specify them according to modeling needs.

1.2.1 Marginal Approaches

Copula Models

Copulas are functions used to construct a joint distribution function or survival function by combining the marginal distributions. Copula models are very well explained in Nelsen (2006) and Joe (1997). A bivariate copula is a function $C(u_1, u_2)$ where $(u_1, u_2) \in [0,1]^2$, with the following properties. The margins of C are uniform: $C(u_1, 1) = u_1$, $C(1, u_2) = u_2$; C is a grounded function: $C(u_1, 0) = C(0, u_2) = 0$ and C is 2-increasing: $C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \geq 0$ for all $(u_1, u_2) \in [0,1]^2$, $(v_1, v_2) \in [0,1]^2$ such that $0 \leq u_1 \leq v_1 \leq 1$ and $0 \leq u_2 \leq v_2 \leq 1$.

Due to Sklar's theorem (Sklar, 1959), if F_1 and F_2 are continuous, then there exists a unique copula C such that for all $t_1, t_2 \geq 0$,

$$F(t_1, t_2) = C(F_1(t_1), F_2(t_2)) \tag{1.8}$$

and if C is a copula and F_1 and F_2 are distribution functions, then the function F in (1.8) is a bivariate distribution function with margins F_1 and F_2.

Sklar's theorem can also be applied to bivariate survivor functions (Georges et al., 2001). Hence if S_1 and S_2 are continuous, then there exists a unique copula $\bar{C}$ such that for all $t_1, t_2 \geq 0$,

$$S(t_1, t_2) = \bar{C}(S_1(t_1), S_2(t_2)) \tag{1.9}$$

where $\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$.

Copulas allow us to separate the marginal distributions from the dependence structure and it is possible to use any marginal survivor functions that are appropriate to the given data such as accelerated failure time or proportional hazards models. If there are any covariates x, $S_j(t_j|x)$, $j = 1, 2$ are modeled and $S(t_1, t_2|x)$ is obtained through (1.9). It is also possible for the copula function to depend on x, but often $C(u_1, u_2)$ is assumed not to depend on x.

Log-Location-Scale (AFT) Models

In log-location-scale models, or accelerated failure time models (AFT), we define $Y_j = \log T_j$, $j = 1, 2$ in the form of

$$Y_{ji} = \mu_j(x) + \sigma_j \epsilon_{ji}, \ i = 1, ..., n \tag{1.10}$$

where $-\infty < y_{ji} < \infty$, $-\infty < \mu_j(x_i) < \infty$, x denotes covariates and $\sigma_j > 0$ for $j = 1, 2$ and $i = 1, ..., n$ so that the bivariate survivor function of (Y_1, Y_2) is written as

$$S(y_1, y_2|x) = S_0\left(\frac{y_1 - \mu_1(x)}{\sigma_1}, \frac{y_2 - \mu_2(x)}{\sigma_2}\right) \tag{1.11}$$

where the survivor function $S_0(\epsilon_1, \epsilon_2)$ for ϵ_1, ϵ_2 does not depend on covariates and $\epsilon_j = \frac{y_j - \mu_j}{\sigma_j}$, $j = 1, 2$.

Bivariate distributions such as bivariate normal, bivariate t or extreme value distributions (Kotz et al., 2000) can be used, or copula models with location-scale marginal survivor functions can be used so that we have the bivariate survivor function

$$S(y_1, y_2|x) = C\left[S_{10}\left(\frac{y_1 - \mu_1(x)}{\sigma_1}\right), S_{20}\left(\frac{y_2 - \mu_2(x)}{\sigma_2}\right)\right]. \tag{1.12}$$

1.2.2 Random Effect Models

A bivariate random effect model assumes that T_1 and T_2 are independent given an unobserved random variable W. Hence, the conditional bivariate survivor function given the random effect is

$$S(t_1, t_2|w) = S_1(t_1|w)S_2(t_2|w) \tag{1.13}$$

and if G is the distribution function of the random variable W, the bivariate survivor function becomes

$$S(t_1, t_2) = \int S_1(t_1|w)S_2(t_2|w)dG(w). \tag{1.14}$$

If there are any covariates x, they can be taken into account by modelling $S_j(t_j|x)$ for $j = 1, 2$ and $S(t_1, t_2|x)$ is obtained through (1.14). Hougaard (2000) contains many details concerning random effects models.

Frailty models

A bivariate frailty model assumes that the conditional survivor functions of T_1 and T_2 given the frailties W_1 and W_2, respectively, are independent and they have conditional proportional hazards model such that

$$S_j(t_j|w_j) = Pr(T_j > t_j|W_j = w_j) = S_{0j}(t_j)^{w_j} \tag{1.15}$$

where $S_{0j}(t_j)$ is a continuous baseline survivor function for $j = 1, 2$. The bivariate survivor function is obtained through (1.14) where G denotes the joint distribution function of (W_1, W_2).

Shared frailty models

A bivariate shared frailty model assumes that T_1 and T_2 are conditionally independent given the frailty W and the lifetime variables satisfy

$$Pr(T_j > t_j|W = w) = S_{0j}(t_j)^w \tag{1.16}$$

where $S_{0j}(t_j)$ is a continuous baseline survivor function for $j = 1, 2$. Hence from (1.14), the unconditional survivor function of (T_1, T_2) is

$$S(t_1, t_2) = \int [S_{01}(t_1)S_{02}(t_2)]^w dG(w) = \varphi^{-1}[-\log S_{01}(t_1) - \log S_{02}(t_2)] \tag{1.17}$$

where $\varphi^{-1}(v) = E[e^{-vW}]$ is the Laplace transform of W. As explained in Section 1.3.1, a class of copulas contain the bivariate frailty models.

Location-scale models

In location-scale models, we have $Y_j = \log T_j$, $j = 1, 2$ in the form of

$$Y_{ji} = w_{ji} + \mu_j(x_i) + \sigma_j \epsilon_{ji} \tag{1.18}$$

where $-\infty < y_{ji} < \infty$, $-\infty < \mu_j(x_i) < \infty$, x_i denotes covariates, $\sigma_j > 0$, (W_{1i}, W_{2i}) has joint distribution function G that does not depend on x_i and $E[W_{ji}] = 0$ for $j = 1, 2$ and $i = 1, ..., n$. The bivariate survivor function of (Y_1, Y_2) is obtained through

$$S(y_1, y_2|x) = \int S_{01}\left(\frac{y_1 - w_1 - \mu_1(x)}{\sigma_1}\right) S_{02}\left(\frac{y_2 - w_2 - \mu_2(x)}{\sigma_2}\right) dG(w_1, w_2). \tag{1.19}$$

Note that it is possible to have $w_{ji} = w_i$ for $j = 1, 2$ and $i = 1, ..., n$ in some situations.

1.2.3 Dependence Measures

It is well-known that the Pearson correlation coefficient effectively measures the linear dependence of two random variables coming from a bivariate normal distribution. However, it may not be a good measure for other bivariate distributions where the conditional mean of Y_i given Y_j is not linear in Y_j. In addition, transformation of Y_1 and Y_2 changes the value of Pearson's correlation. Hence, using a nonparametric correlation type measure which is based on concordance is common. Note that two observations (t_{1i}, t_{2i}) and (t_{1j}, t_{2j}), $i \neq j$ are called concordant if $(t_{1i} - t_{1j})(t_{2i} - t_{2j}) > 0$ and discordant if $(t_{1i} - t_{1j})(t_{2i} - t_{2j}) < 0$.

Two frequently used measures of association based on concordance and discordance are Kendall's tau and Spearman's rho, and their relationship is explained in Nelsen (2006) and Joe (1997). These two measures are defined below, and are invariant with respect to strictly increasing transformations of T_1 and T_2. They equal 1 for the bivariate Fréchet upper bound, $F(t_1, t_2) = \min(F_1(t_1), F_2(t_2))$, and -1 for the Fréchet lower bound, $F(t_1, t_2) = \max(0, F_1(t_1) + F_2(t_2) - 1)$. See Joe (1997, Section 2.2) for a discussion of concordance measures and Joe (1997, Chapter 3) for Fréchet bounds.

Kendall's tau

Kendall's tau is the probability of concordance minus the probability of discordance, that is

$$\begin{aligned}\tau &= Pr((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0) - Pr((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0) \\ &= 4 \int F(t_1, t_2) dF(t_1, t_2) - 1 \\ &= 4E[F(T_1, T_2)] - 1.\end{aligned} \tag{1.20}$$

The range of possible values for τ is $[-1, 1]$.

Spearman's rho

Spearman's rho is defined as the correlation of $F_1(T_1)$ and $F_2(T_2)$ and it is proportional to the probability of concordance minus the probability of discordance for the two vectors (T_{1i}, T_{2i}) and (T_{1j}, T_{2k}), $i \neq j \neq k$ such that the joint distribution function of (T_{1i}, T_{2i}) is $F(t_1, t_2)$ and the joint distribution function of (T_{1j}, T_{2k}) is $F_1(t_1)F_2(t_2)$. Hence Spearman's rho is

$$\begin{aligned}\rho &= 3[Pr((T_{1i} - T_{1j})(T_{2i} - T_{2k}) > 0) - Pr((T_{1i} - T_{1j})(T_{2i} - T_{2k}) < 0)] \\ &= 12 \int \int F_1(t_1) F_2(t_2) dF(t_1, t_2) - 3 \\ &= 12 \int \int S(t_1, t_2) dF_1(t_1) dF_2(t_2) - 3.\end{aligned} \tag{1.21}$$

The range of possible values for ρ is $[-1, 1]$.

Since τ and ρ are invariant to strictly increasing transformations, they can be used as summary measures of dependence for bivariate copulas. The Kendall's tau can be expressed as

$$\begin{aligned}\tau &= 4 \int \int C(u_1, u_2) dC(u_1, u_2) - 1 \\ &= 4E[C(U_1, U_2)] - 1 \\ &= 1 - 4 \int_0^1 \int_0^1 \frac{\partial C(u_1, u_2)}{\partial u_1} \frac{\partial C(u_1, u_2)}{\partial u_2} du_1 du_2\end{aligned} \tag{1.22}$$

and the Spearman's rho becomes

$$\rho = 12 \int \int C(u_1, u_2) du_1 du_2 - 3 = 12 \int \int \bar{C}(u_1, u_2) du_1 du_2 - 3. \tag{1.23}$$

Thus, as is expected and shown in Georges et al. (2001), the Kendall's tau and the Spearman's rho of the survival copula are equal to the Kendall's tau and the Spearman's rho of the associated copula, respectively.

Positive and negative quadrant dependence

Lehmann (1966) introduced the quadrant dependence concept to compare the probability of any quadrant $T_1 \leq t_1, T_2 \leq t_2$ under the distribution function F of (T_1, T_2) with the corresponding probability in the case of independence. (T_1, T_2) is positive quadrant dependent (PQD) if the probability that both T_1 and T_2 large (or small) is not smaller than the probability if they are independent. Hence (T_1, T_2) is PQD if for all $(t_1, t_2) \in \Re^2$

$$Pr(T_1 \leq t_1, T_2 \leq t_2) \geq Pr(T_1 \leq t_1)Pr(T_2 \leq t_2) \tag{1.24}$$

or equivalently,

$$Pr(T_1 \geq t_1, T_2 \geq t_2) \geq Pr(T_1 \geq t_1)Pr(T_2 \geq t_2)$$

or

$$Pr(T_i \geq t_i, T_j \leq t_j) \leq Pr(T_i \geq t_i)Pr(T_j \leq t_j)$$

for $i \neq j$ and $i, j = 1, 2$.

(T_1, T_2) is negative quadrant dependent (NQD) if (1.24) is satisfied for all (t_1, t_2) when the inequality is reversed. Lehmann (1966) showed that if (T_1, T_2) is PQD, then $Cov(T_1, T_2) \geq 0$ if it exists and Kendall's tau and Spearman's rho are nonnegative.

Total positivity of order 2 (TP_2) and reverse rule of order 2 (RR_2)

A bivariate probability density function f is total positivity of order 2 (TP_2) (or, positively likelihood ratio dependent) if

$$f(t_{1i}, t_{2i})f(t_{1j}, t_{2j}) \geq f(t_{1i}, t_{2j})f(t_{1j}, t_{2i}) \tag{1.25}$$

for all $t_{1i} < t_{1j}$, $t_{2i} < t_{2j}$ and $i, j = 1, ..., n$. If f is TP_2 then the distribution function F and the survivor function S are also TP_2. F or S is TP_2 if (1.25) is satisfied when f is replaced by F or S, respectively. Being F or S TP_2 implies F is PQD. Hence, f, F or S being TP_2 is a positive dependence condition.

f is reverse rule of order 2 (RR_2) (or, negatively likelihood ratio dependent) if the inequality in (1.25) is reversed. This is a negative dependence condition.

Tail dependence

Tail dependence measures the dependence between the continuous random variables T_1 and T_2 with distribution functions F_1 and F_2, respectively, in the lower- and upper-quadrant tail. Hence the lower- and upper-tail dependence are defined in Nelsen (2006) as in the following:

The lower tail dependence parameter λ_L is the limit (if it exists) of the conditional probability that T_2 is less than or equal to the $100p^{th}$ percentile of F_2 given that T_1 is less than or equal to the $100p^{th}$ percentile of F_1 as p approaches to 0, i.e.

$$\lim_{p \to 0^+} Pr(T_2 \leq F_2^{-1}(p) | T_1 \leq F_1^{-1}(p)) = \lambda_L$$

and the upper tail dependence parameter λ_U is the limit (if it exists) of the conditional probability that T_2 is greater than the $100p^{th}$ percentile of F_2 given that T_1 is greater than the $100p^{th}$ percentile of F_1 as p approaches to 1, i.e.

$$\lim_{p \to 1^-} Pr(T_2 > F_2^{-1}(p) | T_1 > F_1^{-1}(p)) = \lambda_U.$$

Since copulas are invariant under strictly increasing transformations of the margins, the tail dependences can also be formalized as in Joe (1997, Section 2.1):

If a bivariate copula C is such that

$$\lim_{p \to 0^+} Pr(U_2 \leq p | U_1 \leq p) = \lim_{p \to 0^+} \frac{C(p,p)}{p} = \lambda_L$$

exists, then C has lower tail dependence if $\lambda_L \in (0,1]$ and no lower tail dependence if $\lambda_L = 0$ and if

$$\lim_{p \to 1^-} Pr(U_2 > p | U_1 > p) = \lim_{p \to 1^-} \frac{1 - 2p + C(p,p)}{1-p} = \lambda_U$$

exists, then C has upper tail dependence if $\lambda_U \in (0,1]$ and no upper tail dependence if $\lambda_U = 0$.

Georges et al. (2001) showed that the lower tail dependence of the survival copula is equal to the upper tail dependence of the associated copula and the upper tail dependence of the survival copula is identical to the lower tail dependence of the associated copula.

Local dependence

Clayton (1978) introduced a cross-ratio function that is the ratio of the hazard rate of the conditional distribution of T_1 given $T_2 = t_2$ given in (1.4), to that of T_1 given $T_2 \geq t_2$ given

in (1.5), that is

$$\theta^*(t_1,t_2)=\frac{\lambda_{12}(t_1|t_2)}{\lambda'_{12}(t_1|t_2)}=\frac{S(t_1,t_2)\frac{\partial^2 S(t_1,t_2)}{\partial t_1\partial t_2}}{\frac{\partial S(t_1,t_2)}{\partial t_1}\frac{\partial S(t_1,t_2)}{\partial t_2}}. \tag{1.26}$$

1.3 Copula Models

Copula models have some attractive properties such that (1) the marginal distributions can come from different families; (2) the dependence structure can be investigated separately from the marginal effects since the measures of association do not appear in the marginal distributions; (3) copulas are invariant under strictly increasing transformations of the margins. That is, for G_1 and G_2 monotonic functions, $W_1 = G_1(T_1)$ and $W_2 = G_2(T_2)$ have the same copula as T_1 and T_2. In this section some important parametric families of bivariate copulas are introduced and their properties are given.

The three most important radially symmetric copulas, i.e. $\bar{C} = C$, are the independent copula $C^{\perp}(u_1,u_2) = u_1u_2$, the upper Fréchet copula $C^{+}(u_1,u_2) = \min(u_1,u_2)$ and the lower Fréchet copula $C^{-}(u_1,u_2) = \max(0,u_1+u_2-1)$. Fréchet (1951) indicated that any copula C satisfies the Fréchet-Hoeffding bounds inequality that is $C^{-}(u_1,u_2) \le C(u_1,u_2) \le C^{+}(u_1,u_2)$.

1.3.1 One-Parameter Copula Models

Some well-known bivariate copula classes and their important properties are listed below:

Archimedean copulas

Copulas are called Archimedean when they are of the form

$$C(u_1,u_2)=\varphi^{-1}[\varphi(u_1)+\varphi(u_2)] \tag{1.27}$$

where φ is a decreasing convex function on (0,1] satisfying $\varphi(1) = 0$. The most important characteristic of bivariate Archimedean copulas is that all the information about the 2-dimensional dependence structure is contained in a univariate generator, φ. Some fundamental properties of Archimedean copulas are given in Genest and MacKay (1986), Joe (1997, Section 4.2) and Nelsen (2006, Section 4.3). Archimedean copulas contain the bivariate frailty models when φ^{-1} is the Laplace transform of the underlying frailty distribution (Oakes, 1989).

Some frequently used one-parameter Archimedean families are as follows:

i) Clayton family (Clayton, 1978) has the form

$$C_\phi(u_1, u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, \quad \phi > 0. \tag{1.28}$$

Its generator function is

$$\varphi_\phi(t) = t^{-\phi} - 1 \tag{1.29}$$

and since $\varphi_\phi^{-1}(v) = (v+1)^{-1/\phi}$ is the Laplace trasform of a gamma distribution with index $1/\phi$, bivariate gamma shared frailty model leads to the Clayton survivor copula model.

The Kendall's tau is

$$\tau_\phi = \frac{\phi}{\phi + 2}. \tag{1.30}$$

U_1 and U_2 are positively associated when $\phi > 0$ and the dependence increases as the value of the parameter ϕ increases. The independent copula is obtained when $\phi \to 0$ and the Fréchet upper bound is obtained as $\phi \to \infty$. The lower tail dependence parameter is $\lambda_L = 2^{-1/\phi}$ and it has no upper tail dependence.

ii) Gumbel-Hougaard family (Gumbel, 1960) has the form

$$C_\theta(u_1, u_2) = \exp\left(-\left[(-\log u_1)^\theta + (-\log u_2)^\theta\right]^{1/\theta}\right), \quad \theta > 1. \tag{1.31}$$

Its generator function is

$$\varphi_\theta(t) = (-\log t)^\theta \tag{1.32}$$

and since $\varphi_\theta^{-1}(v) = \exp(-v^{1/\theta})$ is the Laplace transform of a positive stable distribution, positive stable shared frailty model leads to the Gumbel-Hougaard survivor copula model.

The Kendall's tau is

$$\tau_\theta = \frac{\theta - 1}{\theta}. \tag{1.33}$$

The dependence increases as the value of the parameter θ increases. The independent copula is obtained as $\theta \to 1$ and the Fréchet upper bound is obtained as $\theta \to \infty$. The upper tail dependence parameter is $\lambda_U = 2 - 2^{1/\theta}$ and it has no lower tail dependence.

Gumbel-Hougaard copulas are the only Archimedean extreme value copulas, which are defined below (Genest and Rivest, 1989).

iii) Frank family (Frank, 1979) has the form

$$C_\nu(u_1, u_2) = -\frac{1}{\nu}\log\left[1 - \frac{(1 - e^{-\nu u_1})(1 - e^{-\nu u_2})}{1 - e^{-\nu}}\right], \quad \nu \in (-\infty, 0) \cup (0, \infty). \tag{1.34}$$

Its generator function is

$$\varphi_\nu(t) = -\log\left[\frac{1 - e^{-\nu t}}{1 - e^{-\nu}}\right] \tag{1.35}$$

and the Kendall's tau is

$$\tau_\nu = 1 + 4\frac{D_1(\nu) - 1}{\nu} \tag{1.36}$$

where D_1 is the first Debye function, $D_1(\nu) = \int_0^\nu \frac{t}{\nu(e^t-1)}dt$.

U_1 and U_2 are positively associated when $\nu > 0$ and negatively associated when $\nu < 0$. The independent copula is obtained as $\nu \to 0$, the Fréchet upper bound is obtained as $\nu \to \infty$ and the Fréchet lower bound is obtained as $\nu \to -\infty$. It has no lower and upper tail dependence.

Frank copulas are the only Archimedean copulas that satisfy radial symmetry (Frank, 1979).

Extreme-value copulas

Bivariate extreme-value copulas have the form

$$C(u_1, u_2) = \exp\left[\log(u_1u_2)A\left(\frac{\log u_1}{\log(u_1u_2)}\right)\right] \tag{1.37}$$

where A is a convex dependence function on $[0, 1]$ satisfying $\max(t, 1-t) \leq A(t) \leq 1$ for all $t \in [0, 1]$.

The Kendall's tau is (Ghoudi et al., 1998)

$$\tau_A = \int_0^1 \frac{t(1-t)}{A(t)}dA'(t). \tag{1.38}$$

Some common one-parameter extreme-value copulas are given below:

i) Gumbel-Hougaard family which is an Archimedean copula.

ii) Galambos family which has the form

$$C_\phi(u_1, u_2) = u_1u_2\exp[\left((-\log u_1)^{-\phi} + (-\log u_2)^{-\phi}\right)^{-1/\phi}], \quad \phi \geq 0. \tag{1.39}$$

The dependence function is $A_\phi(t) = 1 - [t^{-\phi} + (1-t)^{-\phi}]^{-1/\phi}$. The dependence increases as the value of the parameter ϕ increases. The independent copula is obtained as $\phi \to 0$ and the Fréchet upper bound is obtained as $\phi \to \infty$.

1.3.2 Two or More-Parameter Copula Models

Two or more-parameter copula families provide flexibility for fitting data since they can capture more than one type of dependence. When such a family includes some of the well-known one-parameter copula families such as Clayton, Frank and Gumbel-Hougaard, testing of those models can easily be performed. We provide a pair of examples, and introduce some other models later.

Archimedean copulas

A bivariate two-parameter family of the form of an Archimedean copula is

$$C_{\phi,\theta}(u_1, u_2) = \left\{ \left[(u_1^{-\phi} - 1)^\theta + (u_2^{-\phi} - 1)^\theta \right]^{1/\theta} + 1 \right\}^{-1/\phi}, \tag{1.40}$$

$\phi > 0$ and $\theta \geq 1$. This family includes the Clayton in (1.28) and Gumbel-Hougaard families in (1.31) as special cases. In particular, (1.40) reduces to the Clayton family when $\theta = 1$ and the Gumbel-Hougaard family as $\phi \to 0$. The generator function is

$$\psi_{\phi,\theta}(t) = (t^{-\phi} - 1)^\theta \tag{1.41}$$

and the Kendall's tau is

$$\tau_{\phi,\theta} = 1 - \frac{2}{\theta(\phi + 2)}. \tag{1.42}$$

The dependence increases as the parameters θ and ϕ increase. The independent copula $u_1 u_2$ is obtained as $\phi \to 0$ and $\theta \to 1$ and the Fréchet upper bound $\min(u_1, u_2)$ is obtained as $\phi \to \infty$ or $\theta \to \infty$. Detailed properties of this family are given in Joe (1997, Section 5.2).

Fréchet copulas

Bivariate Fréchet copulas are constructed as a mixture of the independent copula and Fréchet-Hoeffding upper and lower bounds such that

$$C_{\phi,\theta}(u_1, u_2) = (1 - \theta - \phi)u_1 u_2 + \theta \min(u_1, u_2) + \phi \max(u_1 + u_2 - 1, 0), \tag{1.43}$$

$\theta, \phi \in [0, 1]$ and $\theta + \phi \leq 1$. This is a two-parameter reflection symmetric copula family introduced by Fréchet (1958) and a one-parameter version of Fréchet copulas is obtained when $\phi = 0$. Furthermore, it is clear that when $\theta = 0$ and $\phi = 0$, we get the independent copula; when $\theta = 1$ and $\phi = 0$, the upper Fréchet copula; and when $\theta = 0$ and $\phi = 1$, the lower Fréchet copula is obtained.

The Kendall's tau is given as

$$\tau_{\phi,\theta} = \frac{(\theta - \phi)(\theta + \phi + 2)}{3}. \tag{1.44}$$

1.4 Review of Estimation Methods for Copula Models

1.4.1 Estimation with Parallel Clustered Lifetime Data

There are three approaches to specifying and estimating a copula model: fully parametric, semiparametric and fully nonparametric approaches. Fully parametric estimation can be based on one-stage and two-stage procedures, and we consider it first. Suppose the parametric marginal survival functions of T_1 and T_2 are $S_1(t_1;\beta_1)$ and $S_2(t_2;\beta_2)$, respectively and the parametric bivariate survival function of (T_1, T_2) is $S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2))$. Sometimes the marginal distributions are constrained to be the same but for the exposition below we assume that β_1 and β_2 are distinct.

To start, covariates are assumed not to be present. In one-stage estimation, when analyzing parallel right-censored lifetime data, the maximum likelihood estimates of the unknown parameters of the marginal distributions, β_1, β_2 and the copula model, α are obtained simultaneously by maximizing the likelihood function $L(\beta_1, \beta_2, \alpha)$ in (1.6). If $\ell(\beta_1, \beta_2, \alpha)$ denotes the natural logarithm of $L(\beta_1, \beta_2, \alpha)$, the score equations

$$U_{\beta_i}(\beta_1, \beta_2, \alpha) = \frac{\partial \ell(\beta_1, \beta_2, \alpha)}{\partial \beta_i} = 0, \quad i = 1, 2 \tag{1.45}$$

and

$$U_\alpha(\beta_1, \beta_2, \alpha) = \frac{\partial \ell(\beta_1, \beta_2, \alpha)}{\partial \alpha} = 0 \tag{1.46}$$

are solved simultaneously to get the maximum likelihood estimates $\hat{\theta} = (\hat{\beta}_1^t, \hat{\beta}_2^t, \hat{\alpha}^t)^t$ of $\theta = (\beta_1^t, \beta_2^t, \alpha^t)^t$. Under regularity conditions and assuming that the model is correct, $(\hat{\beta}_1, \hat{\beta}_2, \hat{\alpha})$ are consistent estimators of the true values $(\beta_1, \beta_2, \alpha)$ and

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\alpha}\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\\ \alpha\end{pmatrix}\right] \longrightarrow_d MVN(0, J^{-1}(\beta_1, \beta_2, \alpha)) \tag{1.47}$$

where

$$J(\beta_1, \beta_2, \alpha) = \begin{pmatrix} J_{\beta_1\beta_1} & J_{\beta_1\beta_2} & J_{\beta_1\alpha} \\ J_{\beta_2\beta_1} & J_{\beta_2\beta_2} & J_{\beta_2\alpha} \\ J_{\alpha\beta_1} & J_{\alpha\beta_2} & J_{\alpha\alpha} \end{pmatrix} = E\left[-\frac{1}{n}\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^t}\right]$$

is the Fisher information matrix and it depends on censoring process, covariate distribution and $\theta = (\beta_1^t, \beta_2^t, \alpha^t)^t$.

In the two-stage estimation procedure, in the first stage the parametric marginal survival functions are estimated assuming T_1 and T_2 are independent and in the second stage

the dependence parameter of the copula model is estimated. That is, T_1 and T_2 are assumed to be independent and β_1, β_2 are estimated by maximizing the pseudolikelihood function

$$L^*(\beta_1, \beta_2) = \prod_{i=1}^{n} f_1(t_{1i}; \beta_1)^{\delta_{1i}} S_1(t_{1i}; \beta_1)^{1-\delta_{1i}} f_2(t_{2i}; \beta_2)^{\delta_{2i}} S_2(t_{2i}; \beta_2)^{1-\delta_{2i}}. \tag{1.48}$$

If $\ell^*(\beta_1, \beta_2)$ denotes the natural logarithm of $L^*(\beta_1, \beta_2)$, the score equations

$$U^*_{\beta_i}(\beta_1, \beta_2) = \frac{\partial \ell^*(\beta_1, \beta_2)}{\partial \beta_i} = 0, \quad i = 1, 2 \tag{1.49}$$

are solved to get the pseudo-maximum likelihood estimates $(\tilde{\beta}_1, \tilde{\beta}_2)$ of (β_1, β_2) under the independence assumption of T_1 and T_2. Next, the estimate of the dependence parameter α is found by maximizing the likelihood function $L(\tilde{\beta}_1, \tilde{\beta}_2, \alpha)$ that is the likelihood function in (1.6) where β_1, β_2 are replaced by $\tilde{\beta}_1, \tilde{\beta}_2$. If the natural logarithm of $L(\tilde{\beta}_1, \tilde{\beta}_2, \alpha)$ is denoted by $\ell(\tilde{\beta}_1, \tilde{\beta}_2, \alpha)$, then the pseudoscore equation

$$U_\alpha(\tilde{\beta}_1, \tilde{\beta}_2, \alpha) = \frac{\partial \ell(\tilde{\beta}_1, \tilde{\beta}_2, \alpha)}{\partial \alpha} = 0 \tag{1.50}$$

is solved to get the estimate of α. Let $\tilde{\alpha}$ be the solution.

Hence the estimating equation

$$\begin{pmatrix} U^*_{\beta_1}(\beta_1, \beta_2) \\ U^*_{\beta_2}(\beta_1, \beta_2) \\ U_\alpha(\beta_1, \beta_2, \alpha) \end{pmatrix} = 0$$

is solved to get the estimates $(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha})$ of $(\beta_1, \beta_2, \alpha)$. By following the results on estimating equations given in Appendix A, under regularity conditions and the correctness of the model, we conclude that

$$\sqrt{n} \left[\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\alpha} \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \\ \alpha \end{pmatrix} \right] \longrightarrow_d MVN(0, \Sigma) \tag{1.51}$$

where 0 is a zero vector, $\Sigma = A(\beta_1, \beta_2, \alpha)^{-1} B(\beta_1, \beta_2, \alpha)(A(\beta_1, \beta_2, \alpha)^{-1})^t$, where

$$A(\beta_1, \beta_2, \alpha) = \begin{pmatrix} J^*_{\beta_1\beta_1}(\beta_1, \beta_2) & 0 & 0 \\ 0 & J^*_{\beta_2\beta_2}(\beta_1, \beta_2) & 0 \\ J_{\alpha\beta_1}(\beta_1, \beta_2, \alpha) & J_{\alpha\beta_2}(\beta_1, \beta_2, \alpha) & J_{\alpha\alpha}(\beta_1, \beta_2, \alpha) \end{pmatrix},$$

$$B(\beta_1,\beta_2,\alpha)=\begin{pmatrix} J^*_{\beta_1\beta_1}(\beta_1,\beta_2) & J^*_{\beta_1\beta_2}(\beta_1,\beta_2) & 0 \\ J^*_{\beta_2\beta_1}(\beta_1,\beta_2) & J^*_{\beta_2\beta_2}(\beta_1,\beta_2) & 0 \\ 0 & 0 & J_{\alpha\alpha}(\beta_1,\beta_2,\alpha) \end{pmatrix},$$

$$J^*(\beta_1,\beta_2)=\begin{pmatrix} J^*_{\beta_1\beta_1} & J^*_{\beta_1\beta_2} \\ J^*_{\beta_2\beta_1} & J^*_{\beta_2\beta_2} \end{pmatrix}=E\left[-\frac{1}{n}\frac{\partial^2\ell^*(\beta)}{\partial\beta\partial\beta^t}\right]$$

which depends on censoring process, covariate distribution and $\beta=(\beta_1^t,\beta_2^t)^t$.

The details of the above derivations are given in Shih and Louis (1995). Joe (2005) showed that as the underlying copula model approaches the independent copula $C^{\perp}$, the covariance matrix for the estimators obtained through two-stage parametric estimation approaches the covariance matrix for the maximum likelihood estimators. Except when T_1 and T_2 are independent, the two-stage estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are asymptotically less efficient then the one-stage maximum likelihood estimators.

Some authors such as Clayton (1978), Oakes (1982, 1986), Clayton and Cuzick (1985), Hougaard (1989) and Oakes (1994) have used semiparametric two-stage estimation procedures. Genest et al. (1995), Shih and Louis (1995) and Wang and Ding (2000) investigated the asymptotic distribution of the pseudo-maximum likelihood estimator $\tilde{\alpha}$, defined below, under uncensored, right-censored and current status data, respectively.

In the case with no covariates and a parametric copula function $C_\alpha(u_1,u_2)$, the marginal survivor functions are estimated by a nonparametric method in the first stage, giving the nonparametric estimates $\hat{S}_1$ and $\hat{S}_2$ of S_1 and S_2, respectively. Under no censoring, Genest et al. (1995) estimated the marginal survivor functions by the empirical survivor functions and under censoring, Shih and Louis (1995) used Kaplan-Meier estimators of the marginal survivor functions. When the data may be censored, in the second stage the dependence parameter α of the copula model is estimated by maximizing the pseudolikelihood function

$$\begin{aligned} L_s(\alpha)=\prod_{i=1}^{n} c_\alpha(\hat{S}_1(t_{1i}),\hat{S}_2(t_{2i}))^{\delta_{1i}\delta_{2i}}\left[-\frac{\partial C_\alpha(\hat{S}_1(t_{1i}),\hat{S}_2(t_{2i}))}{\partial\hat{S}_1(t_{1i})}\right]^{\delta_{1i}(1-\delta_{2i})} \\ \times\left[-\frac{\partial C_\alpha(\hat{S}_1(t_{1i}),\hat{S}_2(t_{2i}))}{\partial\hat{S}_2(t_{2i})}\right]^{(1-\delta_{1i})\delta_{2i}} C_\alpha(\hat{S}_1(t_{1i}),\hat{S}_2(t_{2i}))^{(1-\delta_{1i})(1-\delta_{2i})} \end{aligned} \tag{1.52}$$

where $c_\alpha(u_1,u_2)=\frac{\partial^2 C_\alpha(u_1,u_2)}{\partial u_1\partial u_2}$ is the copula density function, so that $c_\alpha(S_1(t_1),S_2(t_2))=\frac{\partial^2 C_\alpha(S_1(t_1),S_2(t_2))}{\partial S_1(t_1)\partial S_2(t_2)}$. Note that (1.52) is equivalent to (1.6) with the marginal distributions fixed at their nonparametric estimates. Hence the semiparametric estimate $\tilde{\alpha}$ of α is obtained by solving the estimating equation

$$U_s(\alpha)=\frac{\partial\ell_s(\alpha)}{\partial\alpha}=0 \tag{1.53}$$

where $\ell_s(\alpha) = \log L_s(\alpha)$. Shih and Louis (1995) showed that $\sqrt{n}(\tilde{\alpha} - \alpha)$ converges in distribution to normal with mean zero and a specified variance under some regularity conditions.

Nonparametric copula estimation has also been considered under no censoring; Rüschendorf (1976) and Deheuvels (1979) introduced the empirical copula as

$$C_n\left(\frac{j}{n}, \frac{k}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} I[t_{1i} \leq t_{1(j)}, t_{2i} \leq t_{2(k)}] \tag{1.54}$$

where $t_{1(1)} \leq ... \leq t_{1(n)}$ and $t_{2(1)} \leq ... \leq t_{2(n)}$ are ordered observations of $\{(T_{1i}, T_{2i}), i = 1, ..., n\}$. The empirical copula frequency is $c_n\left(\frac{j}{n}, \frac{k}{n}\right) = \frac{1}{n}$ if $(t_{1(j)}, t_{2(k)}) \in \{(t_{1i}, t_{2i}), i = 1, ..., n\}$, and 0 otherwise. Gijbels and Mielniczuk (1990) and Chen and Huang (2007) proposed kernel estimators of the copula. Chen and Huang (2007) form their nonparametric copula estimator in two stages. First, kernel estimators of the marginal distribution functions $F_1(t_1)$ and $F_2(t_2)$ are found. In the second stage, a kernel copula estimator is obtained based on local linear kernels and a simple mathematical correction that removes the boundary bias. These are not of direct use in this thesis, since we consider censored data.

Under the assumption that the true model is in the class of Archimedean copula models, Genest and Rivest (1993) proposed a nonparametric method for estimating the dependence function C in (1.27) based on uncensored data and Wang and Wells (2000a) suggested a nonparametric estimation procedure for censored data. These methods are summarized in Section 1.5.

Nonparametric estimators of a bivariate survival function based on censored data have been proposed by many authors. Kalbfleisch and Prentice (2002, Section 10.3) survey this area, but see also Campbell (1981), Campbell and Földes (1982), Tsai et al. (1986), Burke (1988), Dabrowska (1988), Pruitt (1990, 1991), Lin and Ying (1993), Prentice and Cai (1992), van der Laan (1996), Wang and Wells (1997), Prentice et al. (2004). Note that in this case, no use of copula models is made.

If there are covariates x, the marginal distributions can be in the form of accelerated failure time, proportional hazards or other regression models. In this case there is no change in fully parametric estimation other than the form of the marginal survival functions. However, if the proportional hazard margins are specified semiparametrically as in the Cox model, then there is so far no complete asymptotic theory for the semiparametric estimation of copula models. He and Lawless (2003) considered instead flexible parametric specifications, e.g. piecewise-constant or spline function specifications, of baseline hazard functions and used one-stage maximum likelihood estimation to fit copula models.

1.4.2 Estimation with Sequential Lifetime Data

Fully parametric estimation for bivariate copula models can be performed by one-stage maximum likelihood estimation in which the likelihood function given in (1.7) is used. Standard two-stage estimation is generally not valid, however, because the pseudolikelihood method in the second stage is not valid when the censoring time for T_2 is not independent of T_2.

The semiparametric estimation procedure for copula models described in Shih and Louis (1995) similarly cannot be used for sequential data because the marginal distribution of T_2 cannot be estimated by the Kaplan-Meier method. He and Lawless (2003) handled the problems in sequential lifetime data by fitting parametric copula models in which the marginal distributions for each of T_1 and T_2 have weakly parametric spline or piecewise-constant forms. One-stage maximum likelihood estimation with the likelihood in (1.7) was used to fit the models.

The nonparametric estimation methods for parallel bivariate data also cannot be applied directly for sequential data because they assume censoring times for T_1, T_2 are independent of (T_1, T_2), whereas this is not true in this situation for T_2. Nonparametric estimation of marginal survivor functions for T_2 were discussed by Visser (1996), Wang and Wells (1998), Wang and Chang (1999), Lin et al. (1999), Van der Laan et al. (2002) and Schaubel and Cai (2004a, 2004b). We describe the approaches of Lin et al. and Schaubel and Cai.

If T_1 is censored then T_2 is not seen. This creates limitations on nonparametric estimation. It is essential in practice to restrict attention to (t_1, t_2) with $t_1 + t_2 \in [0, Q]$ where $Q < C_{\max}$ is a designated value and $C_{\max} = \max(C_1, ..., C_n)$ is the maximum followup time across all subjects. As far as T_2 is concerned, this implies that we can only estimate $Pr(T_2 \leq t_2 | T_1 \leq t_1)$ for similarly restricted (t_1, t_2) values. That is, we can estimate $Pr(T_2 \leq t_2 | T_1 \leq t_1)$ for $0 \leq t_2 \leq C_{\max} - t_1$. Lin et al. (1999) and Schaubel and Cai (2004a) provide nonparametric estimation of $Pr(T_2 \leq t_2 | T_1 \leq t_1)$.

Lin et al. (1999) found an unbiased estimator of the bivariate semi-survival function $H(t_1, t_2) = Pr(T_1 \leq t_1, T_2 > t_2)$ that is given by

$$\tilde{H}(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{I[T_{1i} \leq t_1, T_{2i} > t_2, C_i > T_{1i} + t_2]}{G(T_{1i} + t_2)} \tag{1.55}$$

for values (t_1, t_2) satisfying $t_1 + t_2 \leq C_{\max}$ where $G(c) = Pr(C_i > c)$ is the survivor function of the censoring times C_i. Since G is usually unknown, it is estimated by a Kaplan-Meier estimator $\tilde{G}$ based on the data $\{(t_{1i}, 1 - \delta_{1i}), i = 1, ..., n\}$ or $\{(\min(T_{1i} + T_{2i}, C_i), 1 - \delta_{2i}), i =$

$1, ..., n\}$. Note that an estimator of the marginal distribution of T_1 is obtained as

$$\tilde{F}_1(t_1) = \tilde{H}(t_1, 0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i} I[t_{1i} \leq t_1]}{\tilde{G}(t_{1i})}, \tag{1.56}$$

although this is not necessary because $F_1(t_1)$ can be estimated directly by Kaplan-Meier. In fact, $1 - \tilde{F}_1(t_1)$ is equal to the Kaplan-Meier estimate of $S_1(t_1)$ if $\tilde{G}$ is the Kaplan-Meier estimator of G calculated from $\{(t_{1i}, 1 - \delta_{1i}), i = 1, ..., n\}$. If T_1 has finite support $(0, \tau_1)$ with $\tau_1 < C_{\max}$, $F_2(t_2)$ or $S_2(t_2)$ is estimable for t_2 satisfying $C_{\max} > \tau_1 + t_2$. The estimate of $S_2(t_2)$ is

$$\tilde{H}(\infty, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{I[T_{2i} > t_2, C_i > T_{1i} + t_2]}{\tilde{G}(t_{1i} + t_2)} \tag{1.57}$$

if $Pr(T_1 > C_{\max} - t_2, T_2 > t_2) = 0$. Otherwise, (1.57) is an estimator of $H(C_{\max} - t_2, t_2) = Pr(T_1 \leq C_{\max} - t_2, T_2 > t_2)$. In general, the conditional probability $Pr(T_2 > t_2 | T_1 \leq t_1)$ can be estimated as

$$\tilde{Pr}(T_2 > t_2 | T_1 \leq t_1) = \frac{\tilde{H}(t_1, t_2)}{\tilde{H}(t_1, 0)} \tag{1.58}$$

for (t_1, t_2) satisfying $t_1 + t_2 < C_{\max}$. Since this estimate may not be strictly monotonic in t_2 especially for small sample sizes, Lin et al. (1999) provided a monotonic estimate, that is

$$\tilde{Pr}(T_2 > t_2 | T_1 \leq t_1) = \frac{\min_{s \leq t_2} \tilde{H}(t_1, s)}{\tilde{H}(t_1, 0)}. \tag{1.59}$$

They also provided asymptotic properties of these nonparametric estimates.

Schaubel and Cai (2004a) proposed an estimator of the conditional probability $Pr(T_2 > t_2 | T_1 \leq t_1)$ based on an adjusted version of the Nelson-Aalen cumulative hazard estimator. Similar to Lin et al. (1999), they used the inverse weighting technique when obtaining their estimator, which is given by

$$\tilde{Pr}(T_2 > t_2 | T_1 \leq t_1) = \exp(-\tilde{\Lambda}^*_{21}(t_2 | t_1)) \tag{1.60}$$

where

$$\tilde{\Lambda}^*_{21}(t_2 | t_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{I[t_{1i} \leq t_1, t_{2i} \leq t_2, \delta_{2i} = 1] / \tilde{G}(t_{2i} + t_{1i})}{\frac{1}{n}\sum_{l=1}^{n} \left\{ I[t_{1l} \leq t_1, t_{2l} \geq t_{2i}, \delta_{1i} = 1] / \tilde{G}(t_{2i} + t_{1l}) \right\}}$$

for (t_1, t_2) satisfying $t_1 + t_2 \leq C_{\max}$. G, the survivor function of the censoring times C_i, is estimated by its Kaplan-Meier estimator $\tilde{G}$ based on the data $\{(\min(T_{1i} + T_{2i}, C_i), 1 - \delta_{2i}),$ $i = 1, ..., n\}$. They provided asymptotic properties of the estimator (1.60). In addition,

they estimate the survivor function of T_1 by using the Nelson-Aalen cumulative hazard estimator,

$$\tilde{\Lambda}_1(t_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{I[t_{1i} \leq t_1, \delta_{1i} = 1]}{\frac{1}{n}\sum_{l=1}^{n} I[t_{1l} \geq t_{1i}]}, \tag{1.61}$$

giving, $\tilde{S}_1(t_1) = \exp(-\tilde{\Lambda}_1(t_1))$ for $t_1 \leq C_{\max}$. Alternatively, a Kaplan-Meier estimate could be used.

1.5 Review of Copula Model Selection and Goodness of Fit

In this section we consider methods for assessing the adequacy of a copula specification $C_\alpha(u_1, u_2)$. To do this we try to avoid parametric assumptions for F_1 and F_2. We review here the main methods proposed for assessing the fit of a copula in this framework. We will see that, aside from the special case of the Clayton copula, very few methods deal with censored bivariate lifetime data. First, we consider Archimedean copulas.

1.5.1 Archimedean Copulas

Oakes (1989) showed that for Archimedean copulas (1.27) the conditional hazard ratio $\theta^*(t_1, t_2)$ in (1.26) depends on (t_1, t_2) only through $S(t_1, t_2)$ such that

$$\theta^*(t_1, t_2) = \theta(S(t_1, t_2)) \tag{1.62}$$

where $\theta(v) = -v\frac{\varphi''_\alpha(v)}{\varphi'_\alpha(v)}$, $0 < v < 1$. For example, the Clayton family in (1.28) has $\theta(v) = \phi + 1$, a constant, and the Gumbel-Hougaard family in (1.31) has $\theta(v) = 1 + \frac{\theta - 1}{-\log\ v}$ which is a monotone increasing function of $v \in (0, 1)$.

The function $\varphi_\alpha^{-1}(u)$ is uniquely determined up to a scale change in u by $\theta(v)$ since

$$\varphi_\alpha(v) = \int_v^1 \exp\left[\int_z^{1-\kappa} \frac{\theta(y)}{y} dy\right] dz \tag{1.63}$$

for some constant $\kappa > 0$ and any bivariate survival distribution satisfying (1.62) is Archimedean. Furthermore, Oakes (1989) showed that a conditional version of Kendall's tau

$$\begin{aligned}\tau(t_1, t_2) = {} & Pr((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0 | \min(T_{1i}, T_{1j}) = t_1, \min(T_{2i}, T_{2j}) = t_2) \\ & - Pr((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0 | \min(T_{1i}, T_{1j}) = t_1, \min(T_{2i}, T_{2j}) = t_2)\end{aligned} \tag{1.64}$$

is a function of $\theta(S(t_1,t_2))$ such that

$$\tau(t_1,t_2)=\frac{\theta(S(t_1,t_2))-1}{\theta(S(t_1,t_2))+1} \tag{1.65}$$

where

$$\theta(S(t_1,t_2))=\frac{Pr((T_{1i}-T_{1j})(T_{2i}-T_{2j})>0|\min(T_{1i},T_{1j})=t_1,\min(T_{2i},T_{2j})=t_2)}{Pr((T_{1i}-T_{1j})(T_{2i}-T_{2j})<0|\min(T_{1i},T_{1j})=t_1,\min(T_{2i},T_{2j})=t_2)}. \tag{1.66}$$

Since the bivariate survivor function is unknown, Oakes (1989) conditioned on the size R_{ij} of the corresponding bivariate risk set $\{k:T_{1k}\geq\min(T_{1i},T_{1j}),T_{2k}\geq\min(T_{2i},T_{2j})\}$ in (1.66) and obtained

$$\gamma(r)=\frac{Pr((T_{1i}-T_{1j})(T_{2i}-T_{2j})>0|R_{ij}=r)}{Pr((T_{1i}-T_{1j})(T_{2i}-T_{2j})<0|R_{ij}=r)} \tag{1.67}$$

for $r=2,...,n$. According to Oakes, (1.67) is close to (1.66). Hence, he proposed a graphical Archimedean copula selection procedure by plotting $-1/\log(r/n)$ versus $\gamma(r)$ which is obtained by inserting the estimate of the unknown parameter found based on Kendall's tau.

Several authors have proposed tests for the Clayton copula model. Shih (1998) proposed a goodness of fit test procedure based on its constant conditional hazard ratio property. The test compares unweighted and weighted concordance estimators of the association parameter $\theta=\phi+1$ for the model in (1.28) and if (1.28) is true, these estimators converge to the true value of ϕ and their difference should be close to zero. When there is no censoring, the unweighted concordance estimator given in Oakes (1982) is

$$\hat{\theta}=\frac{\sum_{i<j}\Delta_{ij}}{\binom{n}{2}-\sum_{i<j}\Delta_{ij}} \tag{1.68}$$

where Δ_{ij} is 1 if (T_{1i},T_{2i}) and (T_{1j},T_{2j}) are concordant and 0 if (T_{1i},T_{2i}) and (T_{1j},T_{2j}) are discordant. The asymptotic variance of $\hat{\gamma}=\log\hat{\theta}$ is $V(\eta)/n$ (Oakes, 1982 and Genest et al., 2006b) with $\eta=\frac{1}{\theta-1}$ where $V(\eta)=\frac{8(2\eta+1)^4}{(\eta+1)^2}\frac{1}{4\eta^2}hypergeom([1,1,\eta],[2\eta+1,2\eta+1],1)-\frac{4(2\eta+1)^2(17\eta^3+27\eta^2+14\eta+2)}{3\eta^2(\eta+1)^2(3\eta+1)}$ and hypergeom(.) is the generalized hypergeometric function. Oakes (1986) found the weighted concordance estimator of θ as

$$\hat{\theta}_w=\frac{\sum_{i<j}\Delta_{ij}/R_{ij}}{\sum_{i<j}(1-\Delta_{ij})/R_{ij}}. \tag{1.69}$$

The weights $1/R_{ij}$ assign high weights to the late failures and low weights to the early failures. The asymptotic variance of $\hat{\gamma}_w=\log\hat{\theta}_w$ is $V_w(\eta)/n$ (Oakes, 1986) where $V_w(\eta)=2\eta^2+6\eta+5-(\eta+1)^4\left[\psi'\left(\frac{1}{2}+\frac{\eta}{2}\right)-\psi'\left(1+\frac{\eta}{2}\right)\right]$ and $\psi'(.)$ is the trigamma function.

Shih (1998) showed that under the Clayton model $(\hat{\gamma}_w - \hat{\gamma}) \to_p 0$ and $\sqrt{n}(\hat{\gamma}_w - \hat{\gamma}) \to_d N(0, W(\eta))$ where $W(\eta) = V(\eta) + V_w(\eta) - 2H(\eta)$ and Genest et al. (2006b) found the correct expression of $H(\eta)$ as $-\frac{8\eta^3+19\eta^2+15\eta+3}{\eta^2} + 4(\eta+1)(2\eta+1)^2 \sum_{k=0}^{\infty} \frac{k!\Gamma(2\eta)}{(\eta+k)(2\eta+k+1)}$. When the alternative model's conditional hazard ratio is monotone and the association is strong, Shih (1998) claims that the test is powerful.

Under censoring, Shih (1998) introduced the unweighted concordance estimator as

$$\hat{\theta} = \frac{\sum_{i<j} \Delta_{ij} Z_{ij}}{\sum_{i<j} (1 - \Delta_{ij}) Z_{ij}} \tag{1.70}$$

and the weighted concordance estimator as

$$\hat{\theta}_w = \frac{\sum_{i<j} \Delta_{ij} Z_{ij} / \tilde{R}_{ij}}{\sum_{i<j} (1 - \Delta_{ij}) Z_{ij} / \tilde{R}_{ij}} \tag{1.71}$$

where $Z_{ij} = I[\min(T_{1i}, T_{1j}) \leq C_{1i}, C_{1j}; \min(T_{2i}, T_{2j}) \leq C_{2i}, C_{2j}]$, $\tilde{R}_{ij}$ is the size of the set $\{k : t_{1k} \geq \min(T_{1i}, T_{1j}), t_{2k} \geq \min(T_{2i}, T_{2j})\}$ and (C_{1i}, C_{2i}) $(i = 1, ..., n)$ denote the independent and identically distributed bivariate censoring variables. In this case $\sqrt{n}(\hat{\gamma}_w - \hat{\gamma})$ is asymptotically normally distributed with mean 0 and variance $\tilde{W}(n)$ that can be estimated through $\hat{W} = \sum_{i \neq j \neq k} Z_{ij} Z_{ik} \hat{N}_{ij} \hat{N}_{ik} \left[\frac{1}{\tilde{R}_{ij}\hat{d}} - \frac{2(\hat{\theta}_w+1)}{n\hat{\theta}_w\hat{d}^*}\right] \left[\frac{1}{\tilde{R}_{ik}\hat{d}} - \frac{2(\hat{\theta}_w+1)}{n\hat{\theta}_w\hat{d}^*}\right]$ where $\hat{N}_{ij} = \Delta_{ij}(1+\hat{\theta}_w) - \hat{\theta}_w$, $\hat{d}$ is the proportion of double failures and $\hat{d}^* = \sum_{i<j} Z_{ij} / \binom{n}{2}$. We consider this approach in simulations later in the thesis and it is observed that her asymptotic variance formula sometimes gives negative estimates $\hat{W}$ when there is censoring.

Under the assumption that a copula is Archimedean, Genest and Rivest (1993) noted that the problem of identifying the copula is one-dimensional. They also noted that the generator function φ in

$$S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2)) = \varphi_\alpha^{-1}(\varphi_\alpha(S_1(t_1)) + \varphi_\alpha(S_2(t_2))) \tag{1.72}$$

is uniquely determined by the function

$$K(v; \alpha) = Pr(S(T_1, T_2) \leq v) = v - \frac{\varphi_\alpha(v)}{\varphi'_\alpha(v)}, \ 0 < v \leq 1. \tag{1.73}$$

Indeed φ_α is determined by solving the differential equation

$$\lambda(v; \alpha) = v - K(v; \alpha) = \frac{\partial v}{\partial \log \varphi_\alpha(v)} \tag{1.74}$$

which yields

$$\varphi_\alpha(v) = \exp \int_{v_0}^{v} \frac{1}{\lambda(t; \alpha)} dt \tag{1.75}$$

where $0 < v_0 < 1$ is an arbitrary constant. To get an estimate of φ_α, Genest and Rivest first constructed a nonparametric estimator K_n of K for uncensored data based on a decomposition of Kendall's tau statistic. Define

$$V_i = \frac{1}{n-1} \sum_{j=1}^{n} I[T_{1i} < T_{1j} \ , \ T_{2i} < T_{2j}] \tag{1.76}$$

for $1 \leq i \leq n$. A nonparametric estimator of $K(v;\alpha)$ is then

$$K_n(v) = \sum_{i=1}^{n} \delta(v - V_i)/n \tag{1.77}$$

where $\delta(.)$ denotes the distribution function of a point mass at the origin. Then, by (1.74) a nonparametric estimator of $\lambda(v,\alpha)$ can be found as $\lambda_n(v) = v - K_n(v)$ for $0 < v < 1$ and by using equation (1.75), $\varphi_\alpha(v)$ can be estimated. It is shown in Genest and Rivest (1993) that $K_n(v)$ is a consistent estimator of $K(v;\alpha)$ and the asymptotic variance of $\sqrt{n}(K_n(v) - K(v;\alpha))$ is approximated by

$$K(v;\alpha)(1 - K(v;\alpha)) + k(v;\alpha)[k(v;\alpha)R(v) - 2v(1 - K(v;\alpha))] \tag{1.78}$$

where $k(v;\alpha) = \frac{\partial K(v;\alpha)}{\partial v}$ and $R(v) = 2\int_0^1 (1-t)\varphi_\alpha^{-1}[(1+t)\varphi_\alpha(v)]dt - v^2$.

They proposed a graphical Archimedean copula model selection procedure by plotting and comparing the nonparametric estimate $\lambda_n(v)$ of $\lambda(v;\alpha)$ with $\lambda(v,\hat{\alpha})$ for models under consideration, where $\hat{\alpha}$ is the value for which the theoretical value of Kendall's tau is equal to

$$\tau_n = 4\bar{V} - 1 \tag{1.79}$$

where $\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$. This method of moments estimation is based on the relationship

$$\tau_\alpha = 4E[V] - 1 = 4\int_0^1 \lambda(t;\alpha)dt + 1. \tag{1.80}$$

For two or more-parameter copula families, it is necessary to equate as many as the first few moments of the pseudosample $V_1, ..., V_n$ to the corresponding theoretical expressions to find the estimates of the dependence parameters. As proved in Genest and Rivest (1993), $\sqrt{n}(\tau_n - \tau)/4S \rightarrow_D N(0,1)$ where $S^2 = \sum_{i=1}^{n}(V_i + W_i - 2\bar{V})^2/(n-1)$ and $W_i = \frac{1}{n-1}\sum_{j=1}^{n} I[T_{1j} < T_{1i} \ , \ T_{2j} < T_{2i}]$. Hence, for the one-parameter copula families, the standard error of the estimate $\hat{\alpha}$ can be approximated by applying the delta method to the asymptotic variance of τ_n.

If there is censoring, then this method is not applicable. For bivariate censored data, Wang and Wells (2000a) introduced a method to estimate $K(v;\alpha)$ in (1.73) rewritten as

$$K(v;\alpha) = E[I[S(T_1,T_2) \leq v]] = \int_0^\infty \int_0^\infty I[S(T_1,T_2) \leq v]S(dt_1,dt_2). \tag{1.81}$$

By plugging in a nonparametric estimator $\hat{S}$ of $S(t_1, t_2)$ which takes censoring into account, we get a nonparametric estimator of $K(v;\alpha)$ as

$$\hat{K}(v) = \sum_{i=1}^{n}\sum_{j=1}^{n} I[\hat{S}(t_{1(i)}, t_{2(j)}) \leq v]\hat{S}(\Delta t_{1(i)}, \Delta t_{2(j)}) \tag{1.82}$$

where $t_{1(1)} \leq ... \leq t_{1(n)}$ and $t_{2(1)} \leq ... \leq t_{2(n)}$ are ordered observations of $\{(t_{1i}, t_{2i}), i = 1, ..., n\}$ and $\hat{S}(\Delta t_{1(i)}, \Delta t_{2(j)}) = \hat{S}(t_{1(i)}, t_{2(j)}) - \hat{S}(t_{1(i-1)}, t_{2(j)}) - \hat{S}(t_{1(i)}, t_{2(j-1)}) + \hat{S}(t_{1(i-1)}, t_{2(j-1)})$. $\hat{S}$ can be one of the nonparametric estimators referred to in Section 1.4.1 or 1.4.2 according to the assumption that (T_1, T_2) are independent of (C_1, C_2) or not.

Due to the fact that $S(t_1, t_2)$ cannot capture mass outside the support of (t_1, t_2), equivalently, $K(v;\alpha) = 1 - Pr(S(T_1, T_2) > v)$ can be estimated as follows:

$$\tilde{K}(v) = 1 - \sum_{i=1}^{n}\sum_{j=1}^{n} I[\hat{S}(t_{1(i)}, t_{2(j)}) > v]\hat{S}(\Delta t_{1(i)}, \Delta t_{2(j)}). \tag{1.83}$$

The asymptotic behavior of $\tilde{K}$ depends on that of $\hat{S}$ and a bootstrap method is applied to find an asymptotic variance estimate for $\tilde{K}$. After smoothing $\tilde{K}(v)$, we can estimate φ_α by using the equation in (1.75). Since the estimate of φ_α does not have a tractable form in general, it is more convenient to choose an Archimedean copula model which is closest to the empirical one under a previously defined metric. Hence, Wang and Wells (2000a) proposed a goodness of fit statistic

$$S_{\xi n} = \int_{\xi}^{1} [\tilde{K}(v) - K(v;\alpha)]^2 dv, \ \xi > 0 \tag{1.84}$$

and select the model having the minimum $S_{\xi n}$ compared to the other Archimedean copula models under consideration. Here, since α in $K(v;\alpha)$ is unknown, it is suggested to estimate it by using the relationship in (1.80) as described in Genest and Rivest (1993). However, for censored data, it is not generally convenient to use the proposed estimators of τ_α such as in (1.79). An alternative way proposed by Wang and Wells is estimating α by

$$\tilde{\alpha} = argmin_\alpha \int [\tilde{K}(v) - K(v,\alpha)]^2 dv. \tag{1.85}$$

Now, since the asymptotic behavior of $\tilde{\alpha}$ depends on that of $\tilde{K}$, the bootstrap method is again applied to obtain an asymptotic variance estimate for $\tilde{\alpha}$. Similarly, the asymptotic behavior of $S_{\xi n}$ depends on that of $\tilde{K}$ and the estimate of α. The bootstrap method explained in Genest et al. (2006a, 2008) can be used to estimate the variance of $S_{\xi n}$ for uncensored data.

Equivalently, the graphical model selection procedure described by Genest and Rivest (1993) can be applied to figure out the closest fit. Hence, the plot of the nonparametric estimate $\tilde{\lambda}(v)$ with the plots of $\lambda(v;\tilde{\alpha})$ can be compared for models under consideration.

Genest et al. (2006a) proposed two alternative simple goodness of fit statistics independent of extraneous constant ξ in the case of uncensored data:

$$\begin{aligned} S_n &= \int_0^1 n[K_n(v) - K(v;\hat{\alpha})]^2 k(v;\hat{\alpha})dv \\ &= \frac{n}{3} + n\sum_{i=1}^{n-1} K_n^2\left(\frac{i}{n}\right)\left[K\left(\frac{i+1}{n};\hat{\alpha}\right) - K\left(\frac{i}{n};\hat{\alpha}\right)\right] \\ &- n\sum_{i=1}^{n-1} K_n\left(\frac{i}{n}\right)\left[K^2\left(\frac{i+1}{n};\hat{\alpha}\right) - K^2\left(\frac{i}{n};\hat{\alpha}\right)\right] \end{aligned} \tag{1.86}$$

and

$$\begin{aligned} T_n &= \sup_{0\le v\le 1}\left|\sqrt{n}[K_n(v) - K(v;\hat{\alpha})]\right| \\ &= \sqrt{n}\max_{j=0,1;0\le i\le n-1}\left|K_n\left(\frac{i}{n}\right) - K\left(\frac{i+j}{n};\hat{\alpha}\right)\right| \end{aligned} \tag{1.87}$$

where $k(v;\alpha)$ is the density of $K(v;\alpha)$.

The large-sample distribution of S_n and T_n can also be found for some other common copula models than Archimedean copulas, such as bivariate extreme-value copulas given in (1.37) and Fréchet copulas given in (1.43) where $\phi = 0$. Ghoudi et al. (1998) showed that bivariate extreme value copulas are such that

$$K_A(v) = Pr(S(T_1,T_2)\le v) = v - (1-\tau_A)v\log v \tag{1.88}$$

for $v \in (0,1]$ where τ_A is the Kendall's tau given in (1.38). Genest and Rivest (2001) showed that Fréchet copulas given in (1.43) for $\phi = 0$ can be determined by

$$K(v;\theta) = v - v\log v + v\log\frac{4v}{\left\{[\theta^2 + 4v(1-\theta)]^{1/2} + \theta\right\}^2} \tag{1.89}$$

for $v \in (0,1]$.

Large value of S_n or T_n leads to the rejection of the underlying copula model. The bootstrap procedure described in Genest et al. (2006a, 2008) can be used to obtain an approximate p-value for S_n and T_n since their distributions depend on the association parameter α for all n. Note, however, that these methods do not apply with censored data.

Chen and Fan (2007) used the model selection procedure described in Wang and Wells (2000a). They showed that for selection among candidate copula models that might all be misspecified, estimators of the parametric copulas based on minimizing the selection criterion function in (1.85) may be preferred to other estimators. Since the limiting distribution of Wang and Wells' test statistic depends on model under misspecification and the same data are used twice in obtaining (1.84) where $\alpha = \tilde{\alpha}$ found by (1.85), Chen and Fan proposed a test for model selection from a finite number of Archimedean copulas. They suggested a nonparametric bootstrap procedure to approximate the null distribution of their test statistic.

1.5.2 Non-Archimedean Copulas

Rüschendorf (1976) and Fermanian et al. (2004) found that the bivariate empirical copula process $\sqrt{n}(C_n(u_1, u_2) - C(u_1, u_2))$ tends in law to the Gaussian process where $C_n(u_1, u_2)$ is given in (1.54) or in (1.90) for $d = 2$. However, goodness of fit tests based on empirical copula processes are performed by using bootstrapping which is computationally intensive. Hence, Fermanian (2005) introduced two distribution-free goodness-of-fit test statistics for copulas. The first test uses the idea of the simple chi-square test and considers observed and expected frequencies based on a kernel estimate of the copula density and a parametric estimate of the copula density, respectively. In particular, suppose we have an i.i.d. sample of d-dimentional vectors $T_i = (T_{1i}, ..., T_{di})$, $i = 1, ..., n$ and assume there is no censoring. Let $U_i = (F_1(T_{1i}), ..., F_d(T_{di}))$ denote the vector of the true marginal cumulative distribution functions and $U_{ni} = (F_{n1}(T_{1i}), ..., F_{nd}(T_{di}))$ denote the vector of empirical marginal cumulative distribution functions. The empirical copula process is

$$C_n(u) = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{d} I[F_{nk}(T_{ki}) \leq u_k] \tag{1.90}$$

and a kernel estimator of the copula density at point u is

$$c_n(u) = \frac{1}{h^d} \int K\left(\frac{u - v}{h}\right) C_n(dv) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{u - U_{ni}}{h}\right)$$

where K is a d-dimensional kernel with $\int K = 1$ and $h = h(n)$ is a bandwidth sequence with $h(n) > 0$ and $h(n) \to 0$ as $n \to \infty$. Under some fundamental assumptions related to kernel, the bandwidth and the copula density given in Fermanian (2005),

$$\sqrt{nh^d}((c_n - c)(u_1), ..., (c_n - c)(u_m)) \longrightarrow_d MVN(0, \Sigma)$$

for every m and every vector $u_1, ..., u_m$ in $(0, 1)^d$ such that $c(u_k) > 0$ for every k, where Σ is diagonal and its k^{th} diagonal term is $\int K^2 c^2(u_k)$. Furthermore, if $c_\alpha(u)$ is continuously

differentiable with respect to α in a neighborhood of α_0 for every $u \in (0,1)^d$, then

$$S = \frac{nh^d}{\int K^2} \sum_{k=1}^{m} \frac{(c_n(u_k) - c_{\hat{\alpha}}(u_k))^2}{c_{\hat{\alpha}}(u_k)^2}$$

tends in law towards an m-dimensional chi-squared distribution under the null hypothesis. The power of this test depends on the choice of the arbitrary points $u_1, ..., u_m$.

The other test statistic introduced in Fermanian (2005) is based on the proximity between the smoothed copula density and the estimated parametric density. Under the assumptions given in Theorem 3 in Fermanian (2005),

$$T = \frac{n^2h^d(J_n - (nh^d)^{-1}\int K^2(t)(c_{\hat{\alpha}}w)(u-ht)dtdu + (nh)^{-1}\int c_{\hat{\alpha}}w\sum_{k=1}^{d}\int K_k^2)^2}{2\int c_{\hat{\alpha}}^2 w \int \left(\int K(u)K(u+v)du\right)^2 dv}$$

tends in law towards a chi-squared distribution where $J_n = \int (c_n - K_h * c_{\hat{\alpha}})(u)w(u)du$, w is a weight function, $K_h(.) = K(./h)/h^d$ and $a * b$ is the convolution between a and b. We note again that neither of Fermanian's tests applies to censored data.

Andersen et al. (2005) proposed some test statistics to check whether an assumed one-parameter shared frailty model fits parallel bivariate right censored data without covariates. However, the method can also be applied to other one-parameter copula models than Archimedean copulas. First, the proposed copula model $C_\alpha(u_1, u_2)$ is fitted by semi-parametric estimation without covariates, as described in Section 1.4. That is, F_1 and F_2 are first estimated by Kaplan-Meier, and then α is estimated at stage 2. Then, a nonparametric estimator of the bivariate copula function is also obtained and is compared to the model under consideration. To measure the difference between the semi-parametric estimate $C_{\tilde{\alpha}}$ and the nonparametric estimate $\tilde{C}$ of the copula, Andersen et al. proposed a chi-squared type statistic $\sum_{k=1}^{K}(A_k - B_k)^2$ obtained by partitioning the unit square into K parts. This gives $A_k = \tilde{C}(a_{1k}, a_{2k}) - \tilde{C}(b_{1k}, a_{2k}) - \tilde{C}(a_{1k}, b_{2k}) + \tilde{C}(b_{1k}, b_{2k})$ and $B_k = C_{\tilde{\alpha}}(a_{1k}, a_{2k}) - C_{\tilde{\alpha}}(b_{1k}, a_{2k}) - C_{\tilde{\alpha}}(a_{1k}, b_{2k}) + C_{\tilde{\alpha}}(b_{1k}, b_{2k})$ when the k^{th} part is a rectangle, $[a_{1k}, b_{1k}) \times [a_{2k}, b_{2k})$. They also proposed a Kolmogorov-like statistic $\sup \left| C_{\tilde{\alpha}}(u_1, u_2) - \tilde{C}(u_1, u_2) \right|$ and a statistic based on a weighted difference between the two copulas $\sup_{(u_1,u_2)\in[0,1]\times[0,1]} |\int_{u_1}^{1}\int_{u_2}^{1} G(z_1, z_2)(dC_{\tilde{\alpha}}(z_1, z_2) - d\tilde{C}(z_1, z_2))|$ for $G(z_1, z_2)$ a non-negative weight function. They used a modified bootstrap procedure to obtain the p-values of the tests. There are many loose ends to their procedure, however. In particular, it does not easily handle variable censoring times, the frequency properties of the approach are not clear, their bootstrap procedure is questionable and the procedure appears to lack power, especially when T_1 and T_2 are highly correlated. We consider this approach in simulations later in the thesis.

Chen and Huang (2007) suggested a goodness of fit test statistic which is a Cramér-von Mises type test statistic measuring the distance between the estimated copula $C_{\tilde{\alpha}}$ when

there is no assumption about F_1 and F_2 and their kernel estimator of the copula described in Section 1.4.1 for uncensored data without covariates.

Another goodness of fit test is based on Rosenblatt's probability integral transformation (Rosenblatt, 1952), which is a mapping of a d-variate random vector $(T_1, ..., T_d)$ with an absolutely continuous distribution function $F(t_1, ..., t_d)$ into uniformly and mutually independently distributed variables on the d dimensional hypercube, given by

$$G(t_1) = Pr(T_1 \leq t_1) = F_1(t_1)$$

$$G(t_2) = Pr(T_2 \leq t_2 | T_1 = t_1) = F_{2|1}(t_2|t_1)$$

$$\vdots$$

$$G(t_d) = Pr(T_d \leq t_d | T_1 = t_1, ..., T_{d-1} = t_{d-1}) = F_{d|1,...,d-1}(t_d|t_1, ..., t_{d-1}).$$

Note that the random variables $Y_j = G(T_j)$ for $j = 1, ..., d$ are uniformly and independently distributed on $[0, 1]$. Breymann et al. (2003) used this transformation to perform a goodness of fit test for any copula model, based on uncensored data. Let $C(u_1, ..., u_d)$ be the joint distribution of $U_1 = F_1(T_1), ..., U_d = F_d(T_d)$ for $(u_1, ..., u_d) \in [0, 1]^d$. Then, the conditional distribution of U_j, given $U_1, ..., U_{j-1}$ is

$$C_j(u_j|u_1, ..., u_{j-1}) = \frac{\partial^{j-1} C(u_1, ..., u_j, 1, ..., 1)}{\partial u_1 ... \partial u_{j-1}} \div \frac{\partial^{j-1} C(u_1, ..., u_{j-1}, 1, ..., 1)}{\partial u_1 ... \partial u_{j-1}}$$

for $j = 2, ..., d$ and we know that $C(u_1, 1, ..., 1) = u_1$. Thus, we define

$$Y_{ji} = C_j(F_j(T_{ji})|F_1(T_{1i}), ..., F_{j-1}(T_{j-1i}))$$

for $j = 2, ..., d$ and $i = 1, ..., n$. Breymann et al. (2003) defined $S_i = \sum_{j=1}^{d} [\Phi^{-1}(Y_{ji})]^2$ for $i = 1, ..., n$ where Φ denotes the cumulative distribution function of $N(0, 1)$ random variable. Here, $S_1, ..., S_n$ has a chi-squared distribution with d degrees of freedom under the assumption that the marginal distributions are known. Breymann et al. (2003) assumed that the chi-squared distribution will not be significantly affected by the use of empirical distribution functions for the unknown marginal distribution functions. However, Dobrić and Schmid (2007) showed that the distribution of the test statistic greatly differs from chi-squared distribution when the empirical distribution functions are used.

For bivariate censored data, Klugman and Parsa (1999) performed a goodness of fit test by using Rosenblatt's transformation. They used a Pearson chi-squared statistic computed from $Y_{21}, ..., Y_{2n}$ with some simple modifications due to censored observations. As indicated in Genest et al. (2009), the limiting distribution of the Pearson statistic is not chi-squared since the marginal distribution functions are estimated parametrically. Moreover, their method is invalid for arbitrarily censored data.

By using the idea in Klugman and Parsa (1999), Genest et al. (2009) proposed another goodness of fit test for uncensored data. Since the empirical distribution function

$$D_n(u_1, ..., u_d) = \frac{1}{n} \sum_{i=1}^{n} I[Y_{1i} \leq u_1, ..., Y_{di} \leq u_d], \ (u_1, ..., u_d) \in [0, 1]^d$$

should be close to the independent copula $C^{\perp}(u_1, ..., u_d) = u_1...u_d$ under the null hypothesis, they suggested two statistics

$$S_n^{(C)} = \sum_{i=1}^{n} \left[D_n(Y_{1i}, ...Y_{di}) - C^{\perp}(Y_{1i}, ...Y_{di}) \right]^2$$

and

$$S_n^{(B)} = \frac{n}{3^d} - \frac{1}{2^{d-1}} \sum_{i=1}^{n} \prod_{j=1}^{d} (1 - Y_{ji}^2) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \prod_{k=1}^{d} (1 - \max(Y_{ki}, Y_{kj}))$$

for testing goodness of fit. The limiting null distributions are found by a parametric bootstrap procedure. As with the related tests above, there is no suggestion about the order in which conditioning is done and the procedure applies only to uncensored data.

Finally, Chen and Fan (2005) proposed pseudolikelihood ratio tests for selecting semiparametric copula models fitted as in Genest et al. (1995) under the assumption that there are no censored observations. They first consider selection between two copula models, $C_{1\alpha_1}$ and $C_{2\alpha_2}$. Their pseudolikelihood ratio tests allow both competing copula models to be misspecified. For $i = 1, 2$, let

$$\tilde{\alpha}_i = argmax_{\alpha_i} \left[\frac{1}{n} \sum_{k=1}^{n} \log c_{i\alpha_i}(\tilde{F}_1(t_{1k}), ..., \tilde{F}_d(t_{dk})) \right]$$

where $c_{i\alpha_i}$ is the density function of the copula $C_{i\alpha_i}$ and $\tilde{F}_j(t)$ is the empirical distribution function of T_j, and let

$$\alpha_i^* = argmax_{\alpha_i} E_0[\log c_{i\alpha_i}(U_{1k}, ...U_{dk})]$$

where the expectation E_0 is taken with respect to the true distribution $C_0(F_{10}, ..., F_{d0})$. It was shown that the two-step estimator $\tilde{\alpha}$ is a consistent estimator of α^* and the asymptotic distribution of $\sqrt{n}(\tilde{\alpha} - \alpha^*)$ is Normal with mean 0 and a variance given in Chen and Fan (2005). They test

$$H_0 : E_0 \left[\log \frac{c_{2\alpha_2^*}(U_{1k}, ..., U_{dk})}{c_{1\alpha_1^*}(U_{1k}, ..., U_{dk})} \right] \leq 0$$

and copula model 1 is selected if H_0 is true. To test the null hypothesis, it is first necessary to determine whether the two closest parametric copulas to the true copula are equal, i.e. the two models are generalized nested, or not. Hence they suggested to test if

$$\sigma_a^2 = Var_0\left[\log\frac{c_{2\alpha_2^*}(U_{1k},...,U_{dk})}{c_{1\alpha_1^*}(U_{1k},...,U_{dk})}\right] = 0$$

which happens if and only if the two copula models are generalized nested. They used a bootstrap to approximate the null distribution of the test statistic

$$n\hat{\sigma}_a^2 = \sum_{k=1}^{n}\left[\log\frac{c_{2\tilde{\alpha}_2^*}(U_{1k},...,U_{dk})}{c_{1\tilde{\alpha}_1^*}(U_{1k},...,U_{dk})} - \frac{1}{n}\sum_{l=1}^{n}\log\frac{c_{2\tilde{\alpha}_2^*}(U_{1l},...,U_{dl})}{c_{1\tilde{\alpha}_1^*}(U_{1l},...,U_{dl})}\right]^2$$

If the two models are generalized nonnested, then the pseudolikelihood ratio statistic

$$PLR(\tilde{F}_1,...,\tilde{F}_d;\tilde{\alpha}_1,\tilde{\alpha}_2) = \sum_{i=1}^{n}\log\frac{c_{2\tilde{\alpha}_2}(\tilde{F}_1(T_{1i}),...,\tilde{F}_d(T_{di}))}{c_{1\tilde{\alpha}_1}(\tilde{F}_1(T_{1i}),...,\tilde{F}_d(T_{di}))}$$

is asymptotically distributed as Normal under the null hypothesis and if they are generalized nested, then $2PLR(\tilde{F}_1,...,\tilde{F}_d;\tilde{\alpha}_1,\tilde{\alpha}_2)$ has limiting null distribution of a weighted sum of $\chi^2(1)$ random variables.

If there are more than two copula models to be compared, they tested

$$H_0 : \max_{2\le i\le M} E_0\left[\log\frac{c_{i\alpha_i^*}(U_{1k},...,U_{dk})}{c_{1\alpha_1^*}(U_{1k},...,U_{dk})}\right] \le 0$$

where $C_{1\alpha_1}(u_1,...,u_d)$ is the benchmark model. If there is no significant evidence to reject H_0, it means that no candidate copula model is closer to the true model than the benchmark model. The test statistic includes a trimming function to remove the effect of generalized nested candidate models with the benchmark model. They suggested a bootstrap procedure to approximate the asymptotic null distribution of the given test statistic.

We note that these procedures are complex and that they do not handle censoring. Moreover, they focus on selection among competing models and not testing an individual model.

1.5.3 Simulation Procedures to Compute the P-Value

Estimating a p-value for a test of a null hypothesis is an important task, because many test statistics do not have useable approximations, especially in finite samples. The determination of a limiting distribution of a test statistic under a null hypothesis is indeed

difficult for many tests. It is important to note that a nonparametric bootstrap method does not work in most cases, as the null hypothesis is not respected. Thus, we typically must use a parametric bootstrap procedure to simulate from the null model in order to estimate the p-value for a test.

When a goodness-of-fit test is based on bivariate right censored data, there are three main steps in simulating pseudo data from the null model:

1. generating lifetimes (T_1^*, T_2^*) from the estimated null model,
2. generating censoring times (C_1^*, C_2^*) from an estimate of the distribution of censoring times (C_1, C_2) and
3. computing $t_1^* = \min(T_1^*, C_1^*)$, $t_2^* = \min(T_2^*, C_2^*)$, $\delta_1^* = I[T_1^* = t_1^*]$ and $\delta_2^* = I[T_2^* = t_2^*]$.

Given a censored sample $(t_{1i}^*, t_{2i}^*, \delta_{1i}^*, \delta_{2i}^*, i = 1, ..., n)$, estimation techniques and the corresponding goodness of fit procedure can be applied and the test statistic W^* calculated. These steps are repeated B times and the p-value can then be estimated as

$$p = \sum_{b=1}^{B} I[W_b^* \geq W_{observed}]/B$$

where W_b^* is the value of the test statistic from the b^{th} simulated sample and $W_{observed}$ is the value of the test statistic obtained from the original data.

When a full parametric model is assumed, generation of lifetimes T_1^* and T_2^* is straightforward. However, in many cases F_1 and F_2 are estimated nonparametrically. In addition, the censoring time distribution is often estimated nonparametrically. Thus we are forced to adopt a semiparametric procedure to generate the pseudo data. We provide procedures for doing this in specific settings in subsequent chapters. Here we describe briefly procedures that have been used with censored data.

Andersen et al. (2005) computed p-values for their tests by a simulation-based or modified bootstrap procedure; this test was described in Section 1.5.2. Since they used a semiparametric estimation method to fit the copula model under consideration, the bootstrap procedure is not fully parametric. They also incorporated uncertainty in the estimate $\tilde{\alpha}$ of the dependence parameter α. Thus, the above first step was performed by generating (T_1^*, T_2^*) from the null copula model C_α with $\alpha = \alpha^*$, where α^* was generated from the estimated asymptotic distribution of $\tilde{\alpha}$. They generate (T_1^*, T_2^*) via a shared frailty model which corresponds to a specific Archimedean copula model. For the second step, they generated (C_1^*, C_2^*) from Pruitt's estimator (Pruitt, 1990 and 1991) of the bivariate censoring distribution.

Chen and Huang (2007) proposed a semiparametric bootstrap procedure to estimate p-values after conducting their goodness of fit test for uncensored data without covariates described in Section 1.5.2. Note that their test statistic basically compares an estimated parametric copula model $C_{\tilde{\alpha}}$ when there is no assumption about F_1 and F_2 with their nonparametric copula estimator (see Section 1.4.1). In their procedure, for generating data (t_1^*, t_2^*) from the estimated null copula model $C_{\tilde{\alpha}}$, first $\{t_{1i}^*, i = 1, ..., n\}$ are generated from the empirical distribution $\hat{F}_1$ of $\{t_{1i}, i = 1, ..., n\}$ by sampling with replacement and let $u_{1i}^* = \hat{F}_1(t_{1i}^*)$ for $i = 1, ..., n$. Then, u_{2i}^* are generated from the conditional distribution function of U_2 given $U_1 = u_{1i}^*$, given by $\frac{\partial \hat{C}_{\hat{\alpha}}(u_1,u_2)}{\partial u_1}|_{u_1=u_{1i}^*}$ and then $t_{2i}^* = \hat{F}_2^{-1}(u_{2i}^*)$ where $\hat{F}_2$ is the empirical distribution based on $\{T_{2i}, i = 1, ..., n\}$. We propose another semiparametric bootstrap procedure in Chapter 3. Our bootstrap procedure is similar to one used by Hsieh et al. (2008) in a semicompeting risks setting. However, note that two procedures are identical when carrying out a parametric bootstrap procedure as in Chapter 2.

For Wang and Wells' (2000a) proposed test, they suggested a bootstrap method based on generating a sample $V_1^*, ..., V_n^*$ from $K_{\tilde{\alpha}}(v)$ given in (1.73) where $\tilde{\alpha}$ is the value for which the theoretical value of Kendall's tau is equal to $\tilde{\tau}_n = 4\sum_{i=1}^n V_i^*/n - 1$. However, Genest et al. (2006a) showed that the algorithm is invalid and suggested another bootstrap method based on generating a random sample from $C_{\tilde{\alpha}}$ for uncensored data. Then, the estimation technique described in Genest and Rivest (1993) can be used to obtain $\tilde{\alpha}^*$ and the test statistic in (1.86) or (1.87) can be evaluated based on $\{(T_{1i}^*, T_{2i}^*), i = 1, ..., n\}$. If there is no analytical expression for $K(\alpha)$ in (1.73) under the null hypothesis, a double bootstrap method is used. Genest and Rémillard (2008) and Genest et al. (2009) explain the method in detail.

1.6 Outline of Research

As described above, methods for fitting copula models to data are well developed, but there has been little work on tests of fit for copulas when the lifetimes are subject to censoring. In particular, most of the procedures that have been proposed do not deal with censored data, and those that do suffer from limitations. Moreover, nothing has been done for models that involve covariates. In chapters 2 and 3, we study goodness of fit tests that are based on embedding a proposed copula within a larger parametric family. This allows goodness of fit testing with censored data, and irrespective of whether covariates are in the model. Novel features of our treatment include the use of pseudolikelihood as well as likelihood ratio tests, and consideration of both parametric and semiparametric models. When the proposed and the expanded copula models are estimated by maximum likelihood estimation, the likelihood ratio test is used. However, when they are estimated by two-stage pseudolikelihood estimation, the test is a pseudolikelihood ratio test. The

two-stage procedure requires less computation, which is especially attractive when the marginal lifetime distributions are specified nonparametrically or semiparametrically.

Goodness of fit tests for fully parametric models are considered in Chapter 2, and methods for obtaining p-values for both likelihood ratio and pseudolikelihood ratio tests are given. The performance of the tests is shown to be excellent in simulation studies, including when the expanded copula model is misspecified. It is proved that the likelihood ratio test is consistent, even when the expanded model is misspecified. In Chapter 3, we propose a semiparametric maximum likelihood estimation method in which the copula parameter is estimated without assumptions on the marginal distributions. The efficiency of the two-stage semiparametric estimator (Shih and Louis, 1995) of the copula parameter is compared with that of the semiparametric maximum likelihood estimator of it. Semiparametric estimation procedures are also extended to models with proportional hazards margins. Semiparametric likelihood ratio and pseudolikelihood ratio tests are considered to provide goodness of fit tests for a copula model without making parametric assumptions for the marginal distributions. Semiparametric bootstrap procedures are introduced to obtain p-values for tests. In simulation studies both when the expanded copula family is correct and when it is misspecified, it is observed that the semiparametric pseudolikelihood ratio test is almost as powerful as the parametric likelihood ratio and pseudolikelihood ratio tests while achieving robustness to the form of the marginal distributions. We conclude that the approach is broadly applicable, powerful and easily implemented. We apply the methodology to two data sets in each chapter, one involving covariates and the other without.

There are some difficulties, noted in Section 1.4.2, in modeling and analyzing sequential data. When the sequential survival times for a given individual are not independent, the problem of induced dependent censoring arises for the second and subsequent survival times. Non-identifiability of the marginal survival distributions is another issue since they are observable only if preceding survival times for an individual are uncensored. In addition, in some studies, a significant proportion of individuals may never have the first event. Hence, in Chapter 4, we introduce an approach to address these features of sequential data. We model the joint distribution of the successive survival times by using copula functions. Moreover, we propose some new semiparametric estimation methods in which the copula parameter is estimated without parametric assumptions on the marginal distributions. The performance of semiparametric estimation methods is compared with some other estimation methods in simulation studies and shown to be good. The methodology developed is applied to a motivating example involving relapse and survival following colon cancer treatment. Some goodness of fit tests and informal model checking procedures are also shown and applied in the example. Finally, another approach to model sequential data is introduced by using a copula model for the truncated joint distribution of survival times and a possible way to estimate the copula parameter in this model is described. Properties

of this estimation technique will be investigated in a future work.

We summarize our results, note remaining gaps in methodology and discuss other areas for research in Chapter 5.

Chapter 2

Likelihood-Based Tests of Parametric Copula Models for Parallel Lifetimes

Although methods for copula models are well developed, there has been little work on tests of fit for copulas when the lifetimes are subject to censoring. The marginal distributions can be modeled by various parametric and semiparametric approaches (Lawless, 2003), and well established methods of checking such models can be applied in many settings. However, methods of checking the joint distribution or more specifically, the copula function specifying the association between T_1 and T_2, are less developed. In particular, as reviewed in Section 1.5, most of the procedures that have been proposed do not deal with censored data, and those that do suffer from limitations. We also want methods that apply when covariates are present.

We carry out model checking by the well known device of embedding a proposed copula family in an expanded family of copulas. This allows goodness of fit testing with censored data, and irrespective of whether covariates are in the model. If the proposed and the expanded copula families are estimated by maximum likelihood estimation, then we can check the proposed family by using the likelihood ratio test. However, if they are estimated by the two-stage estimation technique, the model checking may be performed by a pseudolikelihood ratio test. This extends to a formal testing framework the practice of comparing maximized log-likelihoods of competing parametric models (e.g. Genest et al., 1998). In the following section, both large sample approximations and simulation methods for obtaining p-values are presented. In Section 2.2 we provide simulation results that indicate the adequacy of this approach, and compare it with methods of Shih (1998) and Andersen et al. (2005). Section 2.3 applies the methodology to two data sets, one involving covariates and the other without. In Section 2.4 the likelihood ratio test is shown to be consistent, even when the expanded model is misspecified.

2.1 Likelihood Ratio and Pseudolikelihood Ratio Statistics

Suppose the parametric marginal survival functions of T_1 and T_2 are $S_1(t_1;\beta_1)$ and $S_2(t_2;\beta_2)$, respectively and to start assume the model has a one-parameter copula model representation $S(t_1,t_2)=C_\alpha(S_1(t_1),S_2(t_2))$ where α denotes a scalar dependence parameter.

When one-stage maximum likelihood estimation is used, the maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}=(\hat{\beta}_1^t,\hat{\beta}_2^t)^t$ are obtained by solving the score equations in (1.45) and (1.46) where $\beta=(\beta_1^t,\beta_2^t)^t$. When $\alpha=\alpha_0$ is fixed, the maximum likelihood estimates $\hat{\beta}(\alpha_0)=(\hat{\beta}_1(\alpha_0)^t,\hat{\beta}_2(\alpha_0)^t)^t$ are obtained by solving (1.45). We know that under the null hypothesis $H_0:\alpha=\alpha_0$ and under the condition that α_0 and the true values of other parameters are not boundary points in the parameter space, the likelihood ratio statistic

$$\Lambda_1(\alpha_0)=2\ell(\hat{\beta},\hat{\alpha})-2\ell(\hat{\beta}(\alpha_0),\alpha_0) \tag{2.1}$$

is asymptotically distributed as $\chi^2_{(1)}$ where $\ell(\beta,\alpha)$ is the natural logarithm of $L(\beta,\alpha)$ in (1.6) or in (1.7) for parallel clustered lifetime data and sequential lifetime data, respectively. When α_0 is a boundary point and the true values of other parameters are not boundary points, $\Lambda_1(\alpha_0)$ in (2.1) has the limiting distribution with $Pr(\Lambda_1(\alpha_0)\leq q)=0.5+0.5Pr(\chi^2_{(1)}\leq q)$ (Self and Liang, 1987).

When the two-stage fully parametric estimation technique for parallel data is used, the maximum likelihood estimate $\tilde{\beta}=(\tilde{\beta}_1^t,\tilde{\beta}_2^t)^t$ is found under the working independence assumption, i.e. by maximizing the likelihood function given in (1.48), in stage 1. Then, $\tilde{\alpha}$ is obtained through solving the score equation given in (1.50) in stage 2. Now, the pseudolikelihood ratio statistic is defined as

$$\Lambda_2(\alpha_0)=2\ell(\tilde{\beta},\tilde{\alpha})-2\ell(\tilde{\beta},\alpha_0) \tag{2.2}$$

where $\ell(\tilde{\beta},\alpha_0)=\ell(\tilde{\beta}(\alpha_0),\alpha_0)$ because fixing $\alpha=\alpha_0$ does not affect the stage 1 estimate of β. Under the condition that α_0 and the true values of other parameters are not boundary points, the following theorem gives the asymptotic distribution of $\Lambda_2(\alpha_0)$ as a special case in Liang and Self (1996).

Theorem 1. *Under the null hypothesis $H_0:\alpha=\alpha_0$, the limiting distribution of $\Lambda_2(\alpha_0)$ is $\lambda\chi^2_{(1)}$ where $\lambda=J_{\alpha\alpha}(\beta_{10},\beta_{20},\alpha_0)\Sigma_{33}$, $J_{\alpha\alpha}(\beta_{10},\beta_{20},\alpha_0)$ is the last diagonal element of $J(\beta_1,\beta_2,\alpha)$ defined in (1.47) evaluated at $(\beta_1,\beta_2,\alpha)=(\beta_{10},\beta_{20},\alpha_0)$ and Σ_{33} is defined in the following proof.*

Proof. Expanding the log-likelihood $\ell(\beta_1,\beta_2,\alpha)$ in a Taylor series around $(\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\alpha})$ and

evaluating it at $\beta_1 = \beta_{10},\ \beta_2 = \beta_{20},\ \alpha = \alpha_0$, we get

$$\ell(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) = \ell(\beta_{10}, \beta_{20}, \alpha_0) - (\beta_{10} - \tilde{\beta}_1)^t U_{\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) - (\beta_{20} - \tilde{\beta}_2)^t U_{\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) + \frac{1}{2} \begin{pmatrix} \beta_{10} - \tilde{\beta}_1 \\ \beta_{20} - \tilde{\beta}_2 \\ \alpha_0 - \tilde{\alpha} \end{pmatrix}^t I(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) \begin{pmatrix} \beta_{10} - \tilde{\beta}_1 \\ \beta_{20} - \tilde{\beta}_2 \\ \alpha_0 - \tilde{\alpha} \end{pmatrix} + o_p(1) \tag{2.3}$$

where

$$I(\beta_1, \beta_2, \alpha) = \begin{pmatrix} I_{\beta_1\beta_1} & I_{\beta_1\beta_2} & I_{\beta_1\alpha} \\ I_{\beta_2\beta_1} & I_{\beta_2\beta_2} & I_{\beta_2\alpha} \\ I_{\alpha\beta_1} & I_{\alpha\beta_2} & I_{\alpha\alpha} \end{pmatrix} = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^t}$$

is the information matrix and $\theta = (\beta_1^t, \beta_2^t, \alpha)^t$.

Then, expanding the log-likelihood $\ell(\beta_1, \beta_2, \alpha)$ in a Taylor series around $(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0)$ and evaluating it at $\beta_1 = \beta_{10},\ \beta_2 = \beta_{20},\ \alpha = \alpha_0$, we get

$$\begin{aligned} \ell(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) =& \ell(\beta_{10}, \beta_{20}, \alpha_0) - (\beta_{10} - \tilde{\beta}_1)^t U_{\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) - (\beta_{20} - \tilde{\beta}_2)^t U_{\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) + \\ & \frac{1}{2}(\beta_{10} - \tilde{\beta}_1)^t I_{\beta_1\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0)(\beta_{10} - \tilde{\beta}_1) + \\ & \frac{1}{2}(\beta_{20} - \tilde{\beta}_2)^t I_{\beta_2\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0)(\beta_{20} - \tilde{\beta}_2) + \\ & (\beta_{10} - \tilde{\beta}_1)^t I_{\beta_1\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0)(\beta_{20} - \tilde{\beta}_2) + o_p(1). \end{aligned} \tag{2.4}$$

Now, also expanding the score function $U_{\beta_1}(\beta_1, \beta_2, \alpha)$ in a Taylor series around $(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha})$ and evaluating it at $\beta_1 = \tilde{\beta}_1,\ \beta_2 = \tilde{\beta}_2,\ \alpha = \alpha_0$, we get

$$U_{\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) - U_{\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) = -I_{\beta_1\alpha}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha})(\alpha_0 - \tilde{\alpha}) + o_p(n^{1/2}) \tag{2.5}$$

and, similarly,

$$U_{\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) - U_{\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) = -I_{\beta_2\alpha}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha})(\alpha_0 - \tilde{\alpha}) + o_p(n^{1/2}). \tag{2.6}$$

By using (2.3)-(2.6), the pseudolikelihood ratio statistic in (2.2) is written as

$$\Lambda_2(\alpha_0) = \begin{pmatrix} \beta_{10} - \tilde{\beta}_1 \\ \beta_{20} - \tilde{\beta}_2 \\ \alpha_0 - \tilde{\alpha} \end{pmatrix}^t D_n \begin{pmatrix} \beta_{10} - \tilde{\beta}_1 \\ \beta_{20} - \tilde{\beta}_2 \\ \alpha_0 - \tilde{\alpha} \end{pmatrix} + o_p(1) \tag{2.7}$$

where

$$D_n = \begin{pmatrix} I_{\beta_1\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) - I_{\beta_1\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) & I_{\beta_1\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) - I_{\beta_1\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) & 0 \\ I_{\beta_2\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) - I_{\beta_2\beta_1}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) & I_{\beta_2\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) - I_{\beta_2\beta_2}(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0) & 0 \\ 0 & 0 & I_{\alpha\alpha}(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}) \end{pmatrix}.$$

Under regularity conditions, $I(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha})/n = J(\beta_{10}, \beta_{20}, \alpha_0) + o_p(1)$ and $I(\tilde{\beta}_1, \tilde{\beta}_2, \alpha_0)/n = J(\beta_{10}, \beta_{20}, \alpha_0) + o_p(1)$.

Now, from (1.51), it is known that

$$\sqrt{n}\begin{pmatrix}\tilde{\beta}_1 - \beta_{10}\\ \tilde{\beta}_2 - \beta_{20}\\ \tilde{\alpha} - \alpha_0\end{pmatrix} \longrightarrow_d N(0, \Sigma) \tag{2.8}$$

where $\Sigma = A(\beta_{10}, \beta_{20}, \alpha_0)^{-1}B(\beta_{10}, \beta_{20}, \alpha_0)(A(\beta_{10}, \beta_{20}, \alpha_0)^{-1})^t$ and the components of

$$\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} & \Sigma_{13}\\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23}\\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33}\end{pmatrix}$$

are as follows:

$\Sigma_{11} = J^{*-1}_{\beta_1\beta_1}$,

$\Sigma_{12} = J^{*-1}_{\beta_1\beta_1} J^{*}_{\beta_1\beta_2} J^{*-1}_{\beta_2\beta_2}$,

$\Sigma_{13} = -\frac{1}{J_{\alpha\alpha}}(J^{*-1}_{\beta_1\beta_1} J_{\beta_1\alpha} + J^{*-1}_{\beta_1\beta_1} J^{*}_{\beta_1\beta_2} J^{*-1}_{\beta_2\beta_2} J_{\beta_2\alpha})$,

$\Sigma_{21} = J^{*-1}_{\beta_2\beta_2} J^{*}_{\beta_2\beta_1} J^{*-1}_{\beta_1\beta_1}$,

$\Sigma_{22} = J^{*-1}_{\beta_2\beta_2}$,

$\Sigma_{23} = -\frac{1}{J_{\alpha\alpha}}(J^{*-1}_{\beta_2\beta_2} J_{\beta_2\alpha} + J^{*-1}_{\beta_2\beta_2} J^{*}_{\beta_2\beta_1} J^{*-1}_{\beta_1\beta_1} J_{\beta_1\alpha})$,

$\Sigma_{31} = -\frac{1}{J_{\alpha\alpha}}(J_{\alpha\beta_1} J^{*-1}_{\beta_1\beta_1} + J_{\alpha\beta_2} J^{*-1}_{\beta_2\beta_2} J^{*}_{\beta_2\beta_1} J^{*-1}_{\beta_1\beta_1})$,

$\Sigma_{32} = -\frac{1}{J_{\alpha\alpha}}(J_{\alpha\beta_2} J^{*-1}_{\beta_2\beta_2} + J_{\alpha\beta_1} J^{*-1}_{\beta_1\beta_1} J^{*}_{\beta_1\beta_2} J^{*-1}_{\beta_2\beta_2})$ and

$\Sigma_{33} = \frac{1}{J_{\alpha\alpha}} + \frac{1}{J^2_{\alpha\alpha}}(J_{\alpha\beta_1} J^{*-1}_{\beta_1\beta_1} J_{\beta_1\alpha} + J_{\alpha\beta_2} J^{*-1}_{\beta_2\beta_2} J_{\beta_2\alpha} + J_{\alpha\beta_1} J^{*-1}_{\beta_1\beta_1} J^{*}_{\beta_1\beta_2} J^{*-1}_{\beta_2\beta_2} J_{\beta_2\alpha} +$
$J_{\alpha\beta_2} J^{*-1}_{\beta_2\beta_2} J^{*}_{\beta_2\beta_1} J^{*-1}_{\beta_1\beta_1} J_{\beta_1\alpha})$

where $J_{\cdot\cdot}$ and $J^{*}_{\cdot\cdot}$ are the components of $J(\beta_1, \beta_2, \alpha)$ defined in (1.47) and $B(\beta_1, \beta_2, \alpha)$ defined in (1.51), respectively, and they depend on censoring process and covariate distribution. Note that Shih and Louis (1995) showed that under some regularity conditions, $\sqrt{n}(\tilde{\alpha} - \alpha_0) \longrightarrow_d N(0, \Sigma_{33})$.

From the results (2.7) and (2.8), we can reach the conclusion that $\Lambda_2(\alpha_0) \longrightarrow_d \lambda\chi^2_{(1)}$ where λ is the eigenvalue of ΣD and $D = \lim_{n\to\infty} E\left[\frac{1}{n}D_n\right]$. □

When α_0 is a boundary point and the true values of other parameters are not boundary points, $\Lambda_2(\alpha_0)$ in (2.2) has the limiting distribution with $Pr(\Lambda_2(\alpha_0) \le q) = 0.5 +$

$0.5Pr(\lambda\chi^2_{(1)} \leq q)$ (Liang and Self, 1996). Note that λ depends not only on unknown parameters β but also on censoring process and covariate distribution.

If we fit an expanded copula model which has more than one parameter as in Section 1.3.2 and which includes the proposed copula model as a specified case, we can test the proposed model by using a likelihood ratio statistic. Now, suppose the expanded copula model has the vector of dependence parameters (α_1, α_2) and the null model is obtained when $\alpha_2 = \alpha_{20}$. Then, under the null hypothesis $\alpha_2 = \alpha_{20}$, the likelihood ratio statistic

$$\Lambda_1(\alpha_{20}) = 2\ell(\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2) - 2\ell(\hat{\beta}(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}), \alpha_{20}) \tag{2.9}$$

is asymptotically distributed as $\chi^2_{(r)}$ where r is the rank of α_2, provided α_{20} and the true values of other parameters are not boundary points. As we discuss in Section 1.3.2, a number of important copula families correspond to α_{20} being a boundary point in an expanded family, however. Self and Liang (1987) presented the asymptotic distribution of a likelihood ratio statistic for a parameter on the boundary of the parameter space, for different cases. If the dimension of α_2 is 1, α_{20} is a boundary point in the parameter space and the true values of other parameters are not boundary points than $\Lambda_1(\alpha_{20})$ in (2.9) has the limiting distribution with $Pr(\Lambda_1(\alpha_{20}) \leq q) = 0.5 + 0.5Pr(\chi^2_{(1)} \leq q)$.

It can be shown (e.g. Cox and Hinkley, 1974, p.317) that under alternative hypotheses where C_{α_1,α_2} is the correct copula family but $\alpha_2 = \alpha_2^* \neq \alpha_{20}$, the power of the likelihood ratio test approaches 1 as $n \to \infty$. When H_0 is not true but the data do not come from the expanded parametric model under consideration, we show that under regularity conditions, the test based on (2.9) is still consistent; that is, it will reject H_0 with probability 1 as $n \to \infty$. This robustness property is proved in Section 2.4.

When the models are fitted by two-stage parametric estimation, the pseudolikelihood ratio statistic

$$\Lambda_2(\alpha_{20}) = 2\ell(\tilde{\beta}, \tilde{\alpha}_1, \tilde{\alpha}_2) - 2\ell(\tilde{\beta}, \tilde{\alpha}_1(\alpha_{20}), \alpha_{20}) \tag{2.10}$$

can be used to test the null hypothesis, where $\tilde{\alpha}_1(\alpha_{20})$ maximizes $L(\tilde{\beta}, \alpha_1, \alpha_{20})$ with respect to α_1. Pseudolikelihood ratio statistics sometimes have a limiting distribution equivalent to a linear combination of $\chi^2_{(1)}$ random variables under the null hypothesis (Liang and Self, 1996). However, this does not hold when a parameter is on the boundary and in any case the limiting distribution involves the unknown parameters. In view of these points, we will use a parametric bootstrap to obtain p-values when (2.10) is used to test the hypothesis $H_0 : \alpha_2 = \alpha_{20}$. The steps in the simulation procedure must ensure that pseudodata are generated under H_0, and are as follows.

Step 1: Generate an independent bootstrap sample of size n from the estimated null copula model $C_{\tilde{\alpha}_1(\alpha_{20}),\alpha_{20}}(S_1(T_1; \tilde{\beta}_1), S_2(T_2; \tilde{\beta}_2))$ and estimated censoring distribution. When censoring exists, this step is constituted by the three steps given in Section 1.5.3.

Step 2: For the bootstrap sample, estimate β, α by two-stage estimation under both the given null model with $\alpha_2 = \alpha_{20}$ and the expanded family where α_1, α_2 are unrestricted.

Step 3: Calculate the bootstrapped counterpart Λ_2^* of the pseudolikelihood ratio statistic Λ_2 given in (2.10).

Step 4: Steps 1 to 3 are repeated B times and the p-value is estimated as the proportion of times that $\Lambda_2^* \geq \Lambda_2^{obs}$, where Λ_2^{obs} is the observed value of $\Lambda_2(\alpha_{20})$ in the original sample.

We remark that likelihood and pseudolikelihood theory (Self and Liang, 1987; Liang and Self, 1996) suggest that when α_2 is scalar, the limiting distribution for $\Lambda_2(\alpha_{20})$ is the same as for a $\lambda\chi^2_{(1)}$ variable or (when α_{20} is a boundary value) for a $0.5 + 0.5\lambda\chi^2_{(1)}$ variable. Simulations we have conducted suggest that $c + (1-c)\lambda\chi^2_{(1)}$ approximations are satisfactory for moderate sample sizes. This allows us to reduce substantially the number of simulations needed to estimate p-values, as follows. Generate B bootstrap samples and test statistics, say Λ^*_{2b}, $b = 1, ..., B$. First, for $w > 0$, if

$$Pr(\Lambda_2(\alpha_{20}) \leq w) = c + (1-c)Pr(\lambda\chi^2_{(1)} \leq w) \tag{2.11}$$

where $c = Pr(\Lambda_2(\alpha_{20}) = 0)$ is a good approximation can be checked by doing a $\chi^2_{(1)}$ quantile-quantile plot of Λ^*_{2b}'s. Let $\Lambda^*_{2(i)}$ be the i^{th} smallest positive value among the Λ^*_{2b}'s. In the quantile-quantile plot, $\Lambda^*_{2(i)}$'s are plotted against $F^{-1}_{\chi^2_{(1)}}\left(\frac{\frac{i}{B+1} - \hat{c}}{1-\hat{c}}\right)$ where $\hat{c} = \hat{Pr}(\Lambda_2(\alpha_{20}) = 0) = \frac{1}{B+1}\sum_{i=1}^{B} I[\Lambda^*_{2i}(\alpha_{20}) = 0]$ and $F^{-1}_{\chi^2_{(1)}}$ represents the quantile function for the $\chi^2_{(1)}$ distribution. Then, if the plot indicates that $\Lambda^*_{2(i)} \approx \lambda F^{-1}_{\chi^2_{(1)}}\left(\frac{\frac{i}{B+1} - \hat{c}}{1-\hat{c}}\right)$ is a good approximation, a line through the origin can be fitted to the plot and the constant λ can be estimated from it. Finally, the p-value can be estimated by plugging the estimates of λ and c into

$$Pr(\Lambda_2(\alpha_{20}) > \Lambda_2^{obs}) = (1-c)Pr(\chi^2_{(1)} > \Lambda_2^{obs}/\lambda). \tag{2.12}$$

Note that the $\lambda\chi^2_{(1)}$ limiting distributions or approximations involve the unknown parameters. Thus, λ has to be estimated. However, estimating it as above may be better and may require less computation than estimating all of the matrices contributing to λ, especially if derivatives of $\ell(\beta, \alpha_1, \alpha_2)$ are messy.

We remark that in settings where the sample size is not very large, the χ^2 approximation for the distribution of the ordinary likelihood ratio statistic (2.9) may not be very accurate, especially if α_{20} is a boundary point. In that case the simulation procedure just described (with maximum likelihood estimates replacing two-stage estimates) can be used to obtain p-values.

2.1.1 An Illustration

To use the approaches here, we need an expanded copula family that includes the model under consideration. Some two or three parameter copula families that include widely used copulas are given in Joe (1997), Genest et al. (1998) and Nelsen (2006).

For testing the Clayton copula in (1.28) or the Gumbel-Hougaard copula in (1.31), one model that we can use is the two-parameter Archimedean copula family given in (1.40). As noted in Section 1.3.2, it reduces to the Clayton family when $\theta = 1$ and Gumbel-Hougaard family as $\phi \to 0$. Hence, for testing the Clayton model (i.e., $H_0 : \theta = 1$), the likelihood ratio statistic $\Lambda_1(1) = 2\ell(\hat{\beta}, \hat{\phi}, \hat{\theta}) - 2\ell(\hat{\beta}(\theta = 1), \hat{\phi}(\theta = 1), \theta = 1)$ has the limiting distribution $P(\Lambda_1(1) \le q) = 0.5 + 0.5P(\chi^2_{(1)} \le q)$, since $\theta = 1$ is a boundary point, provided the other parameters are not boundary points. For testing the Gumbel-Hougaard model (i.e., $H_0 : \phi = 0$), the likelihood ratio statistic $\Lambda_1(0) = 2\ell(\hat{\beta}, \hat{\phi}, \hat{\theta}) - 2\ell(\hat{\beta}(\phi = 0), \phi = 0, \hat{\theta}(\phi = 0))$ has the same limiting distribution provided other parameters than ϕ are not boundary points.

The pseudolikelihood ratio statistic Λ_2 in (2.10) could also be used to test either $\theta = 1$ or $\phi = 0$, with the simulation procedure described above used to obtain p-values. This statistic has the advantage of slightly simpler estimation of parameters, but the disadvantage of not having a useable large sample approximation to get p-values. However, the limiting distribution for Λ_1 may not be accurate for small or moderate sample sizes, so simulation may sometimes be needed for it also. In order to examine and compare the properties of Λ_1 and Λ_2, we conduct simulation studies described in the next section.

2.2 Simulation Study

A simulation study of the likelihood ratio and pseudolikelihood ratio test statistics under null and alternative hypotheses was conducted. For the null hypothesis we consider the Clayton model in (1.28). The two-parameter copula family (1.40) is used as an expanded family. To assess size and power of the various test statistics, we generated 1000 random bivariate failure time samples of size $n = 100$ from members of the true copula family (1.40), with two degrees of association represented by Kendall's tau values of 0.4 and 0.8. The marginal distributions of the failure times are considered as Weibull with a unit scale parameter and shape parameter 2, but for estimation the marginal distributions F_1 and F_2 are considered as different. We considered both uncensored and censored samples. For the censored case, the bivariate censoring times $C_{1i},\ C_{2i}$ were generated independently as in Shih (1998), and following Andersen et al. (2005), censoring times were assumed to come from an exponential distribution. In one scenario, we generated $C_1,\ C_2$ from exponential distributions so that the probability of censoring for each of $T_1,\ T_2$ is 50%. In a second

scenario, $C_{ji} = m_j + e_{ji}$ where m_j is the median of $F_j(t_j)$ and e_{ji} is exponentially distributed for $j = 1, 2$, so that the probability of censoring in each coordinate is about 30%.

We compare empirical type I error and power of the likelihood ratio and the pseudo-likelihood ratio tests with the tests developed by Shih (1998) and Andersen et al. (2005). These tests can be applied to censored bivariate data without covariates and the goodness of fit test by Shih (1998) is only applicable for testing the Clayton model. Andersen's chi-squared type test statistic was used with the unit square divided into four equal parts, as Andersen et al. (2005) did in their simulation study. For their test statistic, we used the bivariate survivor function estimator described in Gentleman and Vandal (2002) to obtain the nonparametric estimate of the copula. Their nonparametric estimation method computes the maximum likelihood estimate of $S(t_1, t_2)$ for bivariate censored data, but it should be noted that this estimate assigns mass to points, lines and rectangular regions in the plane $\{(t_1, t_2) : t_1 \geq 0, t_2 \geq 0\}$ in general, and so some convention is needed to implement the Andersen et al. approach. In fact this is a major drawback of this approach. To be able to carry out the chi-squared type test statistic in Andersen et al. (2005) accurately, in our simulation we considered the second case for generating censoring times. For the likelihood-based and Andersen test statistics, we estimated for each scenario the 5% critical values from 10000 samples of size 100 under the null hypothesis. For the test of Shih (1998), her asymptotic critical values were used. For the uncensored case, the corrected version of the asymptotic variance formula given in Genest et al. (2006b) was used for Shih's test. The empirical power of the tests was obtained under alternatives belonging to the family (1.40) in which $\phi = 0$, that is, for Gumbel-Hougaard alternatives. The value of θ was in this case chosen to give the values $\tau = 0.4$ and 0.8 for Kendall's tau.

The first six lines of Table 2.1 show the empirical type I error for the likelihood ratio test (Λ_1), pseudolikelihood ratio test (Λ_2), Shih's test (S) and Andersen's test (T). For Λ_1, Λ_2 and T, these values should be close to 0.05 to be consistent with a standard error of about 0.007, based on 1000 samples, since the critical values used for the tests are based on 10000 simulated samples. In the rest of Table 2.1, empirical powers of the tests are shown when the alternative hypothesis is the Gumbel-Hougaard model (1.31). The goodness of fit test introduced by Andersen et al. (2005) is inapplicable for heavy censoring, so results are not shown for it for the case of 50% censoring. Results for Shih's test when there is censoring and $\tau = 0.8$ are similarly not shown. This is because it was observed that under censoring and strong association (larger τ) her asymptotic variance formula sometimes gave negative estimates, so there may be an error. Genest et al. (2006b) have corrected the formula for the case of uncensored data. It is observed that the likelihood ratio and pseudolikelihood ratio tests have empirical type I errors close to 0.05 generally and higher powers than the other two tests. The empirical type I error of Shih's test differs somewhat from 0.05, indicating that the asymptotic approximation given by Shih (1998) is not highly accurate when $n = 100$. Even though it's size appears larger than the nominal 0.05, the

n	% Censored	τ	True copula	Λ_1	Λ_2	S	T
100	0	0.4	$\theta = 1$, $\phi = 1.333$	0.043	0.042	0.034	0.063
100	0	0.8	$\theta = 1$, $\phi = 8$	0.048	0.046	0.069	0.051
100	30	0.4	$\theta = 1$, $\phi = 1.333$	0.043	0.043	0.076	0.056
100	30	0.8	$\theta = 1$, $\phi = 8$	0.052	0.051	NA	0.034
100	50	0.4	$\theta = 1$, $\phi = 1.333$	0.043	0.041	0.062	NA
100	50	0.8	$\theta = 1$, $\phi = 8$	0.050	0.050	NA	NA
100	0	0.4	$\phi = 0$, $\theta = 1.667$	0.998	0.998	0.907	0.636
100	0	0.8	$\phi = 0$, $\theta = 5$	1	1	1	0.584
100	30	0.4	$\phi = 0$, $\theta = 1.667$	0.982	0.981	0.837	0.167
100	30	0.8	$\phi = 0$, $\theta = 5$	1	1	NA	0.031
100	50	0.4	$\phi = 0$, $\theta = 1.667$	0.942	0.941	0.727	NA
100	50	0.8	$\phi = 0$, $\theta = 5$	1	0.996	NA	NA

Table 2.1: Proportion of rejections of $H_0 : \theta = 1$ (i.e., Clayton family), under models (1.40).

power of Shih's test is substantially lower than those for the likelihood-based tests, when there is censoring. The power of Andersen's test is very low compared to that of the other tests.

The power of likelihood and pseudolikelihood ratio tests should of course be high when the true alternative copula model is a member of the expanded copula family (1.40). To check the performance of these tests when this is not the case, we suppose the true copula model is a member of the 3-parameter Archimedean copula family given in Genest et al. (1998), with generator function

$$\varphi(v) = \log\left[\frac{1-(1-\gamma)^{\theta}}{1-(1-\gamma v^{\phi})^{\theta}}\right] \tag{2.13}$$

($\phi > 0$, $\theta > 1$, $0 < \gamma < 1$), but not a member of the two-parameter copula family given in (1.40). The model (2.13) includes the Clayton and Gumbel-Hougaard models as special cases, and it reduces to the Frank copula (1.34) when $\phi = 1$, $\gamma = \nu/\theta$ and $\theta \to \infty$. Since the two-parameter copula family (1.40) does not include the Frank copula, we performed the same simulation study as above, assuming the true copula model is the Frank, with two degrees of association represented by Kendall's tau values of 0.4 and 0.8, which correspond to $\nu = 4.16$ and $\nu = 18.2$ in (1.34), respectively.

The empirical powers of the tests are given in Table 2.2 when the true copula model is the Frank (1.34), the null model is the Clayton (1.28) and the misspecified expanded model is (1.40). The power values for the pseudolikelihood ratio test are generally higher

n	% Censored	τ	Λ_1	Λ_2	S	T
100	0	0.4	0.795	0.816	0.855	0.715
100	0	0.8	0.947	0.997	1.00	0.826
100	30	0.4	0.504	0.513	0.519	0.152
100	30	0.8	0.653	0.752	NA	0.046
100	50	0.4	0.310	0.307	0.277	NA
100	50	0.8	0.421	0.483	NA	NA

Table 2.2: Proportion of rejections of $H_0 : \theta = 1$ (i.e., Clayton family) for a test based on (1.40) but with (1.34) the true copula.

than those of the likelihood ratio test and they are very close to those of Shih's test, which is specifically designed for the Clayton model. Andersen's test again has very low power compared to other tests. With the expanded copula misspecified, we see from Table 2.2 that the power for heavily censored samples (30% and 50%) is much reduced relative to the powers seen in Table 2.1. This and the similarly low power of the Shih and Andersen tests indicate there is limited power to detect departures from a bivariate copula model with effective sample sizes of only 50-70, unless one can rely on help from parametric assumptions.

It helps in the interpretation of power results to compare the Frank copula model when $\nu = 4.16$ and $\nu = 18.2$ with the Clayton and two-parameter copula models, in order to assess how far the Frank model is from fits based on the two-parameter family (1.40). Since all three copulas are Archimedean, they can be determined uniquely by the univariate function $K(v) = Pr(C(T_1, T_2) \leq v)$ in (1.73) defined on the unit interval (Genest and Rivest, 1993). In Figure 2.1, plots of $\lambda(v) = v - K(v)$ in (1.74) are given for the Frank, Clayton and two-parameter copula models. For the Clayton family the copula represented is (1.28) with ϕ equal to the average value of $\hat{\phi}$ over the 1000 samples generated from the Frank model, when the Clayton model is fitted. For (1.40) the copula represented has (ϕ, θ) equal to the average of $(\hat{\phi}, \hat{\theta})$ over the same 1000 samples, when (1.40) is fitted. These models are estimates of the best fitting models to the Frank models within each family. In the top plot of Figure 2.1, it is observed that the Frank copula model with $\nu = 4.16$ ($\tau = 0.4$) is slightly different from the two-parameter copula family (1.40) with $\bar{\hat{\theta}} = 1.365$ and $\bar{\hat{\phi}} = 0.360$, and both differ a good deal from the Clayton model. In the second plot, it is seen that the Frank copula model with $\nu = 18.2$ ($\tau = 0.8$) is quite different than the two-parameter copula family (1.40) with $\bar{\hat{\theta}} = 3.145$ and $\bar{\hat{\phi}} = 0.808$ and so there is not a member of (1.40) that closely approximates the Frank copula. Nevertheless, the statistics Λ_1 and Λ_2 based on (1.40) still have good power, because the best approximating member of (1.40) is again quite different than the Clayton model.

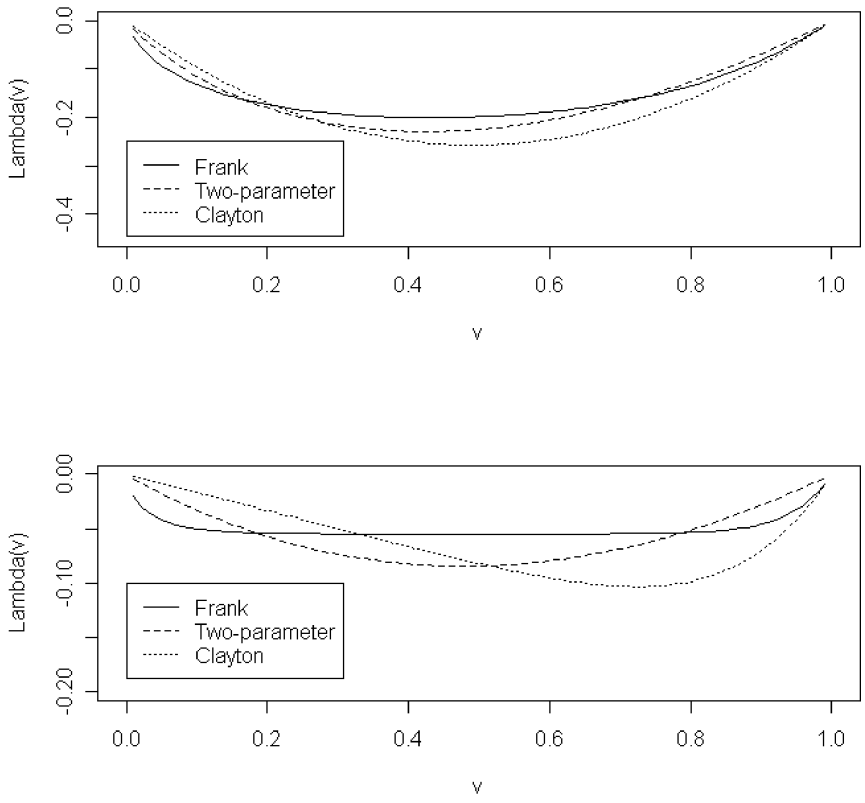


Figure 2.1: Plots of lambda functions for the Frank copula and the best fitting Clayton and two-parameter copula families obtained by the simulation, where the data are generated from the Frank copula with $\nu = 4.16$ (top plot) and $\nu = 18.2$ (bottom plot).

2.3 Applications

2.3.1 Diabetic Retinopathy Study Data

The data described in Section 1.1.1 consists of times or censoring times (t_{1i}, t_{2i}) to the loss of visual acuity in each eye, the treatment indicator $x_{1ji} = I[\text{eye j is treated}]$ for $j = 1, 2$ and type of diabetes indicator $x_{21i} = x_{22i} = I[\text{diabetes is adult-onset}]$ which indicates whether a person's diabetes was adult-onset or juvenile-onset. In total 73% of the lifetimes for treated eyes and 49% of the lifetimes for untreated eyes were censored. Among 197 subjects, 114 had juvenile-onset and 83 had adult-onset diabetes. The primary objective of this study is to understand the effectiveness of the treatment. Since it is paired lifetime data, the independence assumption may be violated.

Huster et al. (1989) considered proportional hazards model

$$h_{ji}(t|x_{ji}) = h_0(t) \exp(\beta_1 x_{1ji} + \beta_2 x_{2ji} + \beta_3 x_{1ji} x_{2ji}) \tag{2.14}$$

for $j = 1, 2$ and $i = 1, ..., 197$ where the baseline hazard function h_0 for the two eyes of a subject or the eyes of any two subjects is the same, and it has Weibull form as the failure time distributions for the eyes. The Clayton copula was used to model the bivariate survival distribution. The treatment, type of diabetes and interaction between them were found to have significant effects on the time to loss of visual acuity. After fitting a working independence proportional hazards model with the same baseline function, it was clear that there is a strong positive association between failure times for the treated and untreated eyes and the treatment effect for adults is more than for juveniles.

Glidden and Self (1999) also considered the Clayton copula with the marginal distributions modeled by a semiparametric proportional hazards regression model given in (2.14). They estimated all the model parameters by an approximate maximum likelihood approach. The same conclusions as in Huster et al. (1989) were reached.

He and Lawless (2003) considered proportional hazards model with the marginal hazards functions in (2.14) for the failure times T_1 and T_2. They employed piecewise constant and spline specifications for the baseline hazard function and the bivariate survival distribution was assumed to be a Clayton copula. Similar results were obtained as in the other papers. He and Lawless (2005) fitted bivariate location-scale models. After the form of the marginal distributions of log-failure-times was assumed to be log-Weibull distribution, the Clayton and Frank bivariate location-scale models were fitted by using a one-stage estimation approach. They also used a bivariate normal model to fit log-failure-times.

Romeo et al. (2006) analyzed the DRS data by copula-based Bayesian parametric and semiparametric estimation procedures. They fitted Clayton, Frank and Gumbel-Hougaard models to the bivariate survival distribution and found in each case the posterior mean

for the dependence parameter under consideration. They also performed a Bayesian semiparametric estimation procedure for each of the specified models above. First the marginal survivor functions were estimated using Kaplan-Meier estimation and then similar to two-stage parametric estimation, the three copula models were estimated by Bayesian posteriors for α. To make the comparison of these models, they used four different approaches: a discrete version of the cross-ratio function defined in Oakes (1989), the predictive model selection approach given in Gelfand and Ghosh (1998), the Bayesian information criterion (BIC) considered in Sahu and Dey (2000) and the average of the logarithm of the pseudo marginal likelihood (ALPML) discussed in Ibrahim et al. (2001). From the comparison of the plots of the posterior cross-ratio functions found from the three Archimedean copula models semiparametrically and from the nonparametric estimate of the cross-ratio function, they decided that the Clayton model (with Kaplan-Meier marginals) is a better fit. When the three copula fits obtained from the two-stage estimation approach with Exponential, Weibull and nonparametric fits of marginals were compared, the predictive model selection approach suggested that the Clayton model with exponential marginals is the best fit. On the other hand, the BIC and ALPML chose the Frank model with Kaplan-Meier marginals as the best fit and the Frank model with Weibull marginals has a better fit than the other models under the parametric marginal distribution assumption. However, when the fits obtained by the one-stage estimation were compared, it was observed that the Clayton model with exponential marginals has the best fit according to the three criteria.

We test here the adequacy of the Clayton and the Gumbel-Hougaard copula families by embedding them in the expanded family of copulas given in (1.40) with $u_1 = S_1(t_1)$, $u_2 = S_2(t_2)$ and the marginal survivor functions modeled by the parametric proportional hazards model in (2.14) with the baseline hazard function of Weibull form, $h_0(t) = \lambda\alpha(\lambda t)^{\alpha-1}$. This has been shown to fit the marginal distributions. In Chapter 3, we also consider semiparametric Cox models as marginal distributions.

The parameters of the proposed and the expanded copula families were estimated by maximum likelihood and the maximum likelihood estimates of the parameters, their standard errors and the maximized log-likelihood values of the corresponding model are given in Table 2.3. From all three models, it is seen that there is a significant treatment effect and the interaction between treatment and type of diabetes has also a significant effect since the treatment is more effective for adult onset diabetics than for the juvenile onset diabetes. A test of the hypothesis $H_0 : \theta = 1$ (i.e., of the Clayton copula) is carried out using (2.9). We obtain the p-value as $0.5P(\chi^2_{(1)} \geq 0.754) = 0.193$, and conclude that there is no evidence against the Clayton model. Similarly, when testing the Gumbel-Hougaard model, $H_0 : \phi = 0$, we obtain the p-value as $0.5P(\chi^2_{(1)} \geq 1.324) = 0.125$. The Clayton model fits a little better since it has a little larger maximum likelihood and a higher p-value, but there is not much difference, and there is no evidence against either model.

The working independence model was also fitted to the data with the marginal dis-

tributions modeled by the parametric proportional hazards model given in (2.14) and the maximum likelihood estimates of the parameters and the maximized log-likelihood value are given in Table 2.4. To test whether there is a significant association between the treated and untreated eyes, i.e. $H_0 : \phi = 0$ in the Clayton model, we can use the likelihood ratio statistic $\Lambda_1(0) = 2(\ell(\hat{\beta}(\theta = 1), \hat{\phi}(\theta = 1), \theta = 1) - \ell(\hat{\beta}(\phi = 0, \theta = 1), \phi = 0, \theta = 1))$ which has a limiting distribution with $Pr(\Lambda_1(0) \leq q) = 0.5 + 0.5Pr(\chi^2_{(1)} \leq q)$. It is obvious that there is a strong positive association between the pair of eyes since $\Lambda_1(0) = 15.886$.

The parameters of the proposed and the expanded copula families were also estimated by two-stage estimation procedure as described in Section 1.4. The estimates of the marginal and dependence parameters, their standard errors and the maximized log-likelihood values are given in Table 2.4. Note that the likelihood ratio statistics are very close to what they were for maximum likelihood estimation. We used the parametric bootstrap method to obtain the p-values. The steps of the procedure are given in Section 2.1 and the first step is explained further for analyzing this data set. In the first step we generate independent samples of size $n = 197$ from the estimated null copula model $C_{\tilde{\eta}_1(\eta_{20}),\eta_{20}}(\tilde{S}_1(T_1), \tilde{S}_2(T_2))$ (i.e. the Clayton model where $\eta_1 = \phi, \eta_{20} = \theta_0 = 1$ or the Gumbel-Hougaard model where $\eta_1 = \theta, \eta_{20} = \phi_0 = 0$) as follows:

1. Generate independent sample $(T^{0*}_{11}, ..., T^{0*}_{1n})$ from the estimated parametric proportional hazards model $\tilde{h}_1(t|x_1) = \tilde{h}_0(t) \exp(\tilde{\beta}_1 x_{11} + \tilde{\beta}_2 x_{21} + \tilde{\beta}_3 x_{11} x_{21})$ where the baseline hazard function $h_0(t)$ is of Weibull form and the covariate vector x_1 is assumed to be fixed at the observed values in the data set.

2. Generate independent sample of size n, $(T^{0*}_{21}, ..., T^{0*}_{2n})$, from $\tilde{S}_{2|1}(T_2|T_1 = T^{0*}_1) = \partial_1 C_{\tilde{\eta}_1(\eta_{20}),\eta_{20}}(\tilde{S}_1(T^{0*}_1), \tilde{S}_2(T_2))$ where $\partial_1 C(u_1, u_2) = \frac{\partial}{\partial u_1} C(u_1, u_2)$, $S_2(t_2) = S_2(t_2|x_2) = [S_0(t_2)]^{\exp(\beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{21} x_{22})}$, $S_0(t_2) = \exp[-H_0(t_2)]$, $H_0(t_2) = \int_0^{t_2} h_0(u)du$ and $h_0(t)$ is of Weibull form.

3. Generate independent sample $(C^*_1, ..., C^*_n)$ from the Kaplan-Meier estimate $\tilde{S}_c$ based on the observed censoring times C_i $(i = 1, ..., m)$ where m is the number of observed censoring times and C_i is observed if and only if $\max(T_{1i}, T_{2i}) > C_i$ for $i = 1, ..., n$. Equivalently, the observed censoring times can be bootstrapped by considering the number of C^*_i's equal to $C_{(j)}$, say N^*_j, where $C_{(j)}$ $(j = 1, ..., k)$ represents the ordered and distinct observed censoring times and k is the number of distinct censoring times. Then, $N^* = (N^*_1, ..., N^*_k)$ has multinomial distribution with sample size n and vector of probabilities $(p_1, ..., p_k)$ where $p_j = \tilde{S}_c(C_{(j-1)}) - \tilde{S}_c(C_{(j)})$ (Efron, 1981 and Reid, 1981).

4. Obtain $(T^*_{j1}, ..., T^*_{jn})$ and $(\delta^*_{j1}, ..., \delta^*_{jn})$ where $T^*_{ji} = \min(T^{0*}_{ji}, C^*_i)$ and $\delta_{ji} = I[T^*_{ji} = T^{0*}_{ji}]$ for $j = 1, 2$ and $i = 1, ..., n$.

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}$	$\hat{\alpha}$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{\beta}_1))$	$(se(\hat{\beta}_2))$	$(se(\hat{\beta}_3))$	$(se(\hat{\lambda}))$	$(se(\hat{\alpha}))$
Two-parameter	-824.880	0.574	1.122	-0.429	0.369	-0.822	0.027	0.811
		(0.549)	(0.149)	(0.182)	(0.198)	(0.297)	(0.007)	(0.062)
Clayton	-825.257	1.006	1	-0.426	0.370	-0.842	0.026	0.818
		(0.331)	(-)	(0.183)	(0.196)	(0.301)	(0.007)	(0.059)
Gumbel-Hougaard	-825.542	0	1.275	-0.429	0.364	-0.793	0.028	0.796
		(-)	(0.093)	(0.183)	(0.198)	(0.294)	(0.008)	(0.063)

Table 2.3: Parametric maximum likelihood estimation results for Diabetic Retinopathy data.

Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\lambda}$	$\tilde{\alpha}$
Working independence	-833.200	0	1	-0.429	0.358	-0.863	0.027	0.813
Two-parameter	-824.907	0.585	1.117					
Clayton	-825.273	1.008	1					
Gumbel-Hougaard	-825.629	0	1.263					

Table 2.4: Two-stage parametric estimation results for Diabetic Retinopathy data.

For each such sample, the two-stage estimation procedure is applied, and produces a value $\Lambda_2^*(\alpha_{20})$. The number of bootstrap samples B was taken to be 1000. When testing the Clayton model, the p-value was estimated as 0.173 and for the Gumbel-Hougaard model, as 0.123. The same conclusion is reached as before, that there is no evidence against either model and the Clayton model fits very slightly better. The p-values are similar to those from the likelihood ratio test.

P-values were also estimated by taking the first $B = 100$ bootstrap samples and using fitted lines to the quantile-quantile plots (see Figure 2.2) as described at the end of Section 2.1. In this case, when testing the Clayton model, the p-value in (2.12) was estimated as 0.175 where the estimates of c and λ were found to be 0.634 and 1.463, respectively. Furthermore, for the Gumbel-Hougaard model, it was estimated as 0.118 where the estimates of c and λ are 0.485 and 0.995, respectively. These estimated p-values are very close to the ones estimated from 1000 bootstrap samples.

2.3.2 Insurance Data

The data described in Frees and Valdez (1998) consist of 1500 general liability claims randomly chosen from claims with late settlement lags. For each claim the indemnity payment (loss) and the allocated loss adjustment expense (ALAE) were recorded. For 1352 claims, the policy limits were also recorded and for the other claims, it is assumed that there are no policy limits. For 34 claims, the amount of the claim is equal to the policy limit, which means they have a censored loss variable. The aim of the study is to fit a joint distribution of losses and expenses. Frees and Valdez (1998), Genest et al. (1998), Klugman and Parsa (1999), Denuit et al. (2004), Chen and Fan (2005) and Genest et al. (2006a) analyzed this data set.

Frees and Valdez (1998) fitted the marginal distributions as Pareto distributions which had been determined by Klugman and Parsa (1995). To identify an appropriate cdf form of the copula, Frees and Valdez used the method developed by Genest and Rivest (1993) assuming the true model is a member of the Archimedean copula family and ignoring the censored observations. They compared the nonparametric estimate of the distribution function $K(v) = Pr(F(X_1, X_2) \leq v)$ where $F(X_1, X_2)$ is the joint distribution function of the loss (X_1) and ALAE (X_2) variables, with three parametric estimates of $K(v)$ corresponding to the Clayton, Frank and Gumbel-Hougaard copulas. The quantile-quantile plots of nonparametric estimate versus parametric estimates of $K(v)$ showed that Gumbel-Hougaard and Frank copula models fit better. After including the censored observations, they fitted these two models by using the one-stage maximum likelihood estimation technique. They compared the Akaike's information criteria (AIC) of the two models and observed that the Gumbel-Hougaard copula model has a better fit.

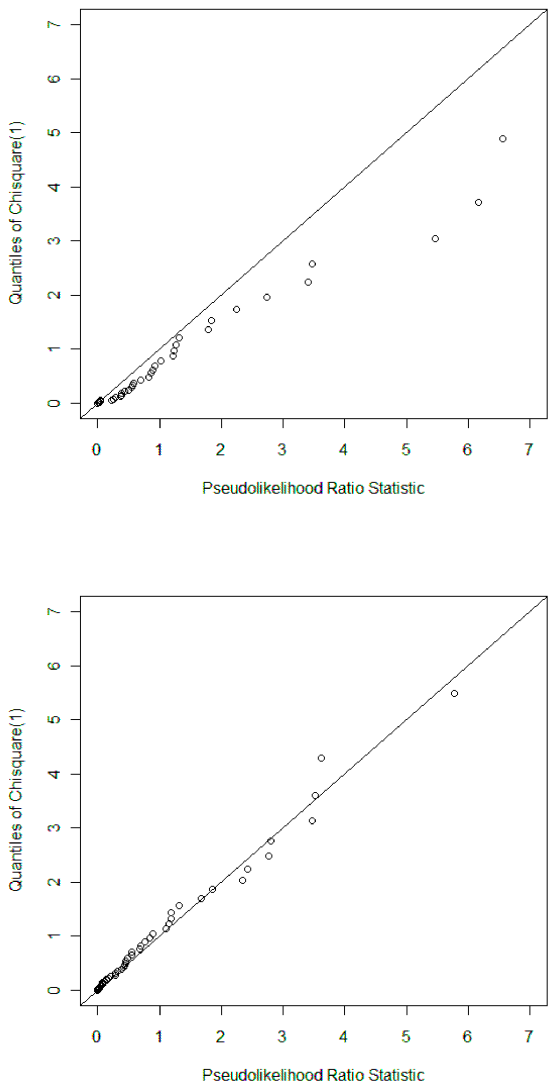


Figure 2.2: $\chi^2_{(1)}$ quantile-quantile plots of 100 simulated values of pseudolikelihood ratio test statistics for testing the Clayton model (top plot) and the Gumbel-Hougaard model (bottom plot).

Genest et al. (1998) used semiparametric estimation as described by Genest et al. (1995) ignoring the censored observations. They used a two-stage pseudolikelihood method where they first estimated the marginal distributions by the empirical distribution function and then fitted four different Archimedean copula models, Clayton, Frank, Gumbel-Hougaard and the three-parameter copula family (2.13), which includes the other three models. By an informal comparison of the pseudolikelihoods, they decided that the Gumbel-Hougaard model is preferred. They also found that the introduction of asymmetry in the Gumbel-Hougaard copula does not improve the fit.

Klugman and Parsa (1999) showed that the inverse paralogistic and the inverse Burr distributions represent the loss and ALAE variables, respectively, better than the Pareto distributions. They fitted the Frank copula model by using the one-stage maximum likelihood estimation approach.

Denuit et al. (2004) followed the procedure presented by Wang and Wells (2000a) under the assumption that the Archimedean copula family includes the true joint distribution of loss and ALAE. The censored observations were taken into account by using the estimator of the bivariate distribution function given in Akritas (1994). They compared the nonparametric estimate of the distribution function $K(v) = Pr(F(X_1, X_2) \leq v)$ with four parametric estimates of $K(v)$ corresponding to the Clayton, Frank, Gumbel-Hougaard and Joe copulas (Joe, 1997, Section 5.1). The quantile-quantile plots of nonparametric estimate versus parametric estimates of $K(v)$ and the plots of nonparametric and parametric estimates of $\lambda(v)$ in (1.74) versus v showed that Gumbel-Hougaard and Frank copula models fit better. However, they selected the former model since it has the minimum L^2-norm distance (1.84) among the four copula models although the latter model has a very close distance value to the former one.

Chen and Fan (2005) used the same estimation technique as in Genest et al. (1998). They fitted Gaussian, Student's t, survival Clayton and a mixture of the Clayton and the Gumbel-Hougaard copulas as well as the copula models that Denuit et al. (2004) fitted. They used pseudolikelihood ratio tests as described in Section 1.5.2 to select the closest model to the true model. They obtained strong evidence that none of the other copulas performs significantly better than the Gumbel-Hougaard copula.

Genest et al. (2006a) compared Clayton, Frank and Gumbel-Hougaard copula models based on S_n, T_n and Wang and Wells' statistic S_{ξ_n} with $\xi = 0$ given in equations (1.86), (1.87) and (1.84), respectively, by ignoring the censored observations. According to all of the three test statistics, they selected the Gumbel-Hougaard copula model.

Here, the forms of the marginal distributions are assumed to be inverse paralogistic and inverse Burr distributions for loss and ALAE variables, respectively. Diagnostic checks show these models to be satisfactory, and they have been used previously by Klugman and

Parsa (1999). For convenience, the cumulative distribution functions

$$F_1(x_1) = \left(\frac{x_1^{\alpha_1}}{x_1^{\alpha_1} + \beta_1^{\alpha_1}}\right)^{\alpha_1}$$

for the inverse paralogistic distribution and

$$F_2(x_2) = \left(\frac{x_2^{\gamma}}{x_2^{\gamma} + \beta_2^{\gamma}}\right)^{\alpha_2}$$

for the inverse Burr distribution are used. The adequacy of the Clayton and Gumbel-Hougaard models is tested by embedding them in (1.40) with $u_1 = F_1(t_1)$ and $u_2 = F_2(t_2)$, since this distribution function copula form has been used by previous authors. First the two-parameter Archimedean copula model in (1.40) and the reduced models in (1.28) and (1.31) are fitted by full maximum likelihood; the results are given in Table 2.5. When testing the Clayton model, $H_0 : \theta = 1$, the likelihood ratio statistic is $\Lambda_1(1) = 212.979$ and there is very strong evidence against the model. However, when testing the Gumbel-Hougaard model, $H_0 : \phi = 0$, it is observed that there is no evidence against the Gumbel-Hougaard model since the likelihood ratio statistic $\Lambda_1(0)$ is approximately 0. Indeed, the value of the statistic is surprisingly small.

To test whether there is a significant association between loss and ALAE variables, i.e. $H_0 : \theta = 1$ in the Gumbel-Hougaard model, the working independence model is fitted (see Table 2.6). It is seen that there is a strong association between the two variables.

Two-stage parametric estimation method and pseudolikelihood ratio test were also applied (see Table 2.6). The values of the pseudolikelihood ratio statistics for testing the Clayton and the Gumbel-Hougaard models are very similar to the likelihood ratio statistics. To get the p-values of the tests, the parametric bootstrap method given in Section 2.1 is applied. The first step is explained further for analyzing this data set. We generate independent samples of size $n = 1500$ from the estimated null copula model $C_{\tilde{\eta}_1(\eta_{20}),\eta_{20}}(\tilde{F}_1(X_1), \tilde{F}_2(X_2))$ (i.e. the Clayton model where $\eta_1 = \phi, \eta_{20} = \theta_0 = 1$ or the Gumbel-Hougaard model where $\eta_1 = \theta, \eta_{20} = \phi_0 = 0$) as follows:

1. Generate independent sample of size n, $(X_{11}^{0*}, ..., X_{1n}^{0*})$, from the estimated inverse paralogistic distribution $\tilde{F}_1(x_1) = (x_1^{\tilde{\alpha}_1}/(x_1^{\tilde{\alpha}_1} + \tilde{\beta}_1^{\tilde{\alpha}_1}))^{\tilde{\alpha}_1}$ under the working independence assumption.

2. Generate independent sample of size n, $(X_{21}^*, ..., X_{2n}^*)$, from $\tilde{F}(X_2|X_1 = X_1^{0*}) = \partial_1 C_{\tilde{\eta}_1(\eta_{20}),\eta_{20}}(\tilde{F}_1(X_1^{0*}), \tilde{F}_2(X_2))$ where $\partial_1 C(u_1, u_2) = \frac{\partial}{\partial u_1} C(u_1, u_2)$ and $\tilde{F}_2(x_2) = (x_2^{\tilde{\gamma}}/(x_2^{\tilde{\gamma}} + \tilde{\beta}_2^{\tilde{\gamma}}))^{\tilde{\alpha}_2}$.

3. Generate independent sample of size n, $(C_{11}^*, ..., C_{1n}^*)$, from the Kaplan-Meier estimate $\tilde{S}_c$ of the observed censoring times corresponding to the loss (X_1) variable.

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$	$\hat{\gamma}$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{\alpha}_1))$	$(se(\hat{\beta}_1))$	$(se(\hat{\alpha}_2))$	$(se(\hat{\beta}_2))$	$(se(\hat{\gamma}))$
Two-parameter	-31741.01	2×10^{-5}	1.445	1.045	11130.51	0.643	8591.020	1.533
		(≈ 0)	(0.032)	(0.017)	(532.619)	(0.050)	(711.693)	(0.060)
Clayton	-31847.50	0.570	1	1.023	11905.79	0.543	10415.55	1.643
		(0.052)	(-)	(0.017)	(589.068)	(0.043)	(823.150)	(0.072)
Gumbel-Hougaard	-31741.01	0	1.445	1.045	11130.98	0.643	8589.908	1.533
		(-)	(0.032)	(0.017)	(532.638)	(0.050)	(711.873)	(0.060)

Table 2.5: Parametric maximum likelihood estimation results for insurance loss data.

Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$	$\tilde{\alpha}_1$	$\tilde{\beta}_1$	$\tilde{\alpha}_2$	$\tilde{\beta}_2$	$\tilde{\gamma}$
Working independence	-31945.70	0	1	1.043	11262.460	0.640	8601.760	1.539
Two-parameter	-31741.10	9×10^{-5}	1.445					
Clayton	-31850.81	0.516	1					
Gumbel-Hougaard	-31741.10	0	1.445					

Table 2.6: Two-stage parametric estimation results for insurance loss data.

4. Obtain $(X_{11}^*, ..., X_{1n}^*)$ and $(\delta_{11}^*, ..., \delta_{1n}^*)$ where $X_{1i}^* = \min(X_{1i}^{0*}, C_{1i}^*)$ and $\delta_{1i} = I[X_{1i}^* = X_{1i}^{0*}]$ for $i = 1, ..., n$.

The number of bootstrap samples B mentioned in the fourth step is taken to be 1000. Since the p-values are very similar to those obtained from likelihood ratio statistics, the same conclusion is reached as before, that there is no evidence against the Gumbel-Hougaard model and very strong evidence against the Clayton model.

2.4 Consistency of Likelihood Ratio Test

In this section, we investigate the performance of the likelihood ratio test when the expanded copula family is misspecified. Suppose the expanded copula family is the two-parameter copula C_{α_1,α_2}, β represents the vector of parameters in the marginal distributions and the parameter of interest is α_2, that is, the null hypothesis under consideration is $H_0 : \alpha_2 = \alpha_{20}$.

First, we show the consistency of the likelihood ratio test when the expanded copula family $F(t_1, t_2; \alpha_1, \alpha_2, \beta)$ is not misspecified. Suppose that $\alpha_2 = \alpha_2^* \neq \alpha_{20}$ and $F(t_1, t_2; \alpha_1, \alpha_{20}, \beta) \neq F(t_1, t_2; \alpha_1, \alpha_2^*, \beta)$. Let $\hat{\theta} = (\hat{\beta}^t, \hat{\alpha}_1, \hat{\alpha}_2)^t$ be the maximum likelihood estimator of $\theta = (\beta^t, \alpha_1, \alpha_2)^t$ and $\theta^* = (\beta^{*t}, \alpha_1^*, \alpha_2^*)^t$ be the true parameter vector. Expanding the log-likelihood $\ell(\theta)$ in a Taylor series around $\hat{\theta}$ and evaluating it at $\theta = \theta^*$, we get

$$\ell(\hat{\theta}) = \ell(\theta^*) + \frac{1}{2}(\theta^* - \hat{\theta})^t I(\hat{\theta})(\theta^* - \hat{\theta}) + o_p(1) \tag{2.15}$$

where $I(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^t}$. Then, expanding the log-likelihood $\ell(\theta)$ in a Taylor series around $\hat{\theta}_0 = (\hat{\beta}(\alpha_{20})^t, \hat{\alpha}_1(\alpha_{20}), \alpha_{20})^t$ and evaluating it at $\theta_0^* = (\beta_0^{*t}, \alpha_{10}^*, \alpha_{20})^t$ where $(\hat{\beta}(\alpha_{20})^t, \hat{\alpha}_1(\alpha_{20}))^t \xrightarrow{p} (\beta_0^{*t}, \alpha_{10}^*)^t$, we get

$$\ell(\hat{\theta}_0) = \ell(\theta_0^*) + \frac{1}{2}(\theta_0^* - \hat{\theta}_0)^t I(\hat{\theta}_0)(\theta_0^* - \hat{\theta}_0) + o_p(1). \tag{2.16}$$

By using (2.15) and (2.16), the likelihood ratio statistic in (2.9) is written as

$$\Lambda_1(\alpha_{20}) = 2(\ell(\theta^*) - \ell(\theta_0^*)) + (\hat{\theta} - \theta^*)^t I(\hat{\theta})(\hat{\theta} - \theta^*) - (\hat{\theta}_0 - \theta_0^*)^t I(\hat{\theta}_0)(\hat{\theta}_0 - \theta_0^*) + o_p(1). \tag{2.17}$$

When the true values of the parameters are not boundary points or when only α_{20} is a boundary point, the second and the third terms in (2.17) are asymptotically distributed as chi-squared with degrees of freedom p and $p-1$, respectively, where p is the total number of parameters. When only α_1^* or only α_2^* is a boundary point, the second term is asymptotically distributed as $0.5\chi_{p-1}^2 + 0.5\chi_p^2$. Asymptotic distributions of likelihood

ratio statistics are given in Self and Liang (1987) under different cases for parameters being boundary points. In any case, under mild conditions asymptotic distributions of the second and the third terms are mixtures of chi-square distributions. The first term is asymptotically positive and unbounded as shown in Cox and Hinkley (1974, pages 288 and 317) and, therefore, $\Lambda_1(\alpha_{20})$ is a consistent test, i.e. for any finite value c, $Pr(\Lambda_1(\alpha_{20}) > c|\alpha_2 = \alpha_2^* \neq \alpha_{20}) \to 1$ as sample size $n \to \infty$.

Theorem 2. *Given the assumptions A1-A6 in White (1982), the likelihood ratio test is a consistent test when the expanded model is misspecified.*

Proof. Suppose the expanded model C_{α_1,α_2} is misspecified and f_θ represents the misspecified distribution. Let g represent the true distribution. White (1982) showed that given the assumptions A1-A3, $\hat{\theta}$ is a consistent estimator of θ^*, where θ^* is the value of θ in the parameter space minimizing the Kullback-Leibler Information Criterion

$$E_G\left[\log\left(\frac{g(T_1,T_2)}{f_\theta(T_1,T_2)}\right)\right] \tag{2.18}$$

uniquely. Similarly, $\hat{\gamma}_0 = (\hat{\beta}(\alpha_{20})^t, \hat{\alpha}_1(\alpha_{20}))^t$ is a consistent estimator of $\gamma_0^* = (\beta_0^{*t}, \alpha_{10}^*)^t$, where γ_0^* is the value of $\gamma = (\beta^t, \alpha_1)^t$ in the parameter space minimizing (2.18) uniquely subject to $\alpha_2 = \alpha_{20}$.

Expanding the log-likelihood $\ell(\theta)$ in a Taylor series around $\hat{\theta}$ and evaluating it at $\theta = \theta^*$, we get

$$\ell(\hat{\theta}) = \ell(\theta^*) + \frac{n}{2}(\theta^* - \hat{\theta})^t A_n(\hat{\theta})(\theta^* - \hat{\theta}) + o_p(1) \tag{2.19}$$

where $A_n(\theta)$ is defined in Appendix A. Then, expanding the log-likelihood $\ell(\theta)$ in a Taylor series around $\hat{\theta}_0$ and evaluating it at θ_0^*, we get

$$\ell(\hat{\theta}_0) = \ell(\theta_0^*) + \frac{n}{2}(\theta_0^* - \hat{\theta}_0)^t A_n(\hat{\theta}_0)(\theta_0^* - \hat{\theta}_0) + o_p(1). \tag{2.20}$$

By using (2.19) and (2.20), the likelihood ratio statistic in (2.9) is written as

$$\Lambda_1(\alpha_{20}) = 2(\ell(\theta^*) - \ell(\theta_0^*)) + n(\hat{\theta} - \theta^*)^t A_n(\hat{\theta})(\hat{\theta} - \theta^*) - n(\hat{\theta}_0 - \theta_0^*)^t A_n(\hat{\theta}_0)(\hat{\theta}_0 - \theta_0^*) + o_p(1). \tag{2.21}$$

White (1982) showed that

$$\sqrt{n}(\hat{\theta} - \theta^*) \to_d N(0, A(\theta^*)^{-1} B(\theta^*)[A(\theta^*)^{-1}]^t) \tag{2.22}$$

as described in Appendix A and by following the notations there. Thus, the second term in (2.21)

$$n(\hat{\theta} - \theta^*)^t A_n(\hat{\theta})(\hat{\theta} - \theta^*) \to_d \sum_{i=1}^{p} \mu_{1i} \chi^2_{(1)} \tag{2.23}$$

where μ_{1i}'s $(i = 1, ..., p)$ are eigenvalues of

$$\lim_{n\to\infty} E_G[A_n(\hat{\theta})]A(\theta^*)^{-1}B(\theta^*)[A(\theta^*)^{-1}]^t = B(\theta^*)[A(\theta^*)^{-1}]^t.$$

Furthermore,

$$\sqrt{n}(\hat{\gamma}_0 - \gamma_0^*) \to_d N(0, [A(\theta_0^*)]^{-1}_{\gamma\gamma}[B(\theta_0^*)]_{\gamma\gamma}([A(\theta_0^*)]^{-1}_{\gamma\gamma})^t) \tag{2.24}$$

where $A(\theta)$ and $B(\theta)$ are partitioned into submatrices such that

$$A(\theta) = \begin{pmatrix} [A(\theta)]_{\gamma\gamma} & [A(\theta)]_{\gamma\alpha_2} \\ [A(\theta)]_{\alpha_2\gamma} & [A(\theta)]_{\alpha_2\alpha_2} \end{pmatrix}, B(\theta) = \begin{pmatrix} [B(\theta)]_{\gamma\gamma} & [B(\theta)]_{\gamma\alpha_2} \\ [B(\theta)]_{\alpha_2\gamma} & [B(\theta)]_{\alpha_2\alpha_2} \end{pmatrix},$$

$[A(\theta)]_{\gamma\gamma}$ and $[B(\theta)]_{\gamma\gamma}$ are $(p-1) \times (p-1)$ matrices. Therefore, the third term in (2.21)

$$n(\hat{\theta}_0 - \theta_0^*)^t A_n(\hat{\theta}_0)(\hat{\theta}_0 - \theta_0^*) = n(\hat{\gamma}_0 - \gamma_0^*)^t [A_n(\hat{\theta}_0)]_{\gamma\gamma}(\hat{\gamma}_0 - \gamma_0^*) \to_d \sum_{i=1}^{p-1} \mu_{0i}\chi^2_{(1)} \tag{2.25}$$

where μ_{0i}'s $(i = 1, ..., p-1)$ are eigenvalues of

$$\lim_{n\to\infty} E_G[[A_n(\hat{\theta}_0)]_{\gamma\gamma}][A(\theta_0^*)]^{-1}_{\gamma\gamma}[B(\theta_0^*)]_{\gamma\gamma}([A(\theta_0^*)]^{-1}_{\gamma\gamma})^t = [B(\theta_0^*)]_{\gamma\gamma}([A(\theta_0^*)]^{-1}_{\gamma\gamma})^t.$$

On the other hand, the first term $2(\ell(\theta^*) - \ell(\theta_0^*))$ is unbounded since the assumption in White (1982) that the Kullback-Leibler Information Criterion (2.18) has a unique minimum at θ^* in the parameter space requires (in the case with no censoring)

$$E_G\left[\log\left(\frac{g(T_1,T_2)}{f_{\theta^*}(T_1,T_2)}\right)\right] < E_G\left[\log\left(\frac{g(T_1,T_2)}{f_{\theta_0^*}(T_1,T_2)}\right)\right] \tag{2.26}$$

if $\alpha_2^* \neq \alpha_{20}$. Then, (2.26) becomes

$$E_G\left[\log f_{\theta^*}(T_1,T_2) - \log f_{\theta_0^*}(T_1,T_2)\right] > 0. \tag{2.27}$$

Let $E_G\left[\log f_{\theta^*}(T_1,T_2)\right] = c^*$ and $E_G\left[\log f_{\theta_0^*}(T_1,T_2)\right] = c_0^*$. By the strong law of large numbers, $\frac{1}{n}\ell(\theta^*) \to c^*$ with probability 1 and $\frac{1}{n}\ell(\theta_0^*) \to c_0^*$ with probability 1. By using (2.27), $\frac{1}{n}\left[\ell(\theta^*) - \ell(\theta_0^*)\right] \to c^* - c_0^* > 0$ with probability 1 and, therefore, the first term in (2.21) is asymptotically positive and unbounded. Hence, when the expanded model is misspecified, for any finite value c, $Pr(\Lambda_1(\alpha_{20}) > c) \to 1$ as sample size $n \to \infty$.

Under mild conditions, this result can be extended to censored data with covariates when f_θ and g are the misspecified and true distributions of $(T_1, T_2, C_1, C_2, x_1, x_2)$, respectively. □

When α_{20} is a boundary point, the results obtained in the above proof are still correct. Furthermore, Theorem 2 is also valid when any parameter values in the vectors θ^* or γ_0^* are on the boundary as the first term in (2.21) is unbounded in any case and the other terms are asymptotically distributed as mixtures of chi-square distributions.

Chapter 3

Estimation and Tests of Fit for Semiparametric Models

An advantage of the approaches in Chapter 2 is that they can also be applied to testing semiparametric copula models in which the marginal distributions are nonparametric or semiparametric. We develop this approach in this chapter. In this case the proposed and the expanded models are fitted by semiparametric estimation using either semiparametric maximum likelihood or two-stage semiparametric estimation. In the following section, we introduce a semiparametric maximum likelihood estimation method. The two-stage semiparametric estimation technique suggested by Shih and Louis (1995) for models without covariates was summarized in Section 1.4.1. In Section 3.1.2 we generalize this method to regression models with Cox proportional hazards margins. Likelihood and pseudolikelihood ratio statistics for testing a semiparametric copula model are given in Section 3.2, and semiparametric bootstrap procedures are suggested to estimate p-values. A major advantage of this approach is that we are not forced to adopt fully parametric assumptions for the marginal distributions. In Section 3.3 efficiency of semiparametric maximum likelihood and two-stage semiparametric estimates of a copula dependence parameter is studied and power comparisons of the likelihood ratio and pseudolikelihood ratio tests are given. Section 3.4 applies the methodology to the two data sets analyzed in the previous chapter.

3.1 Semiparametric Estimation of Copula Models

3.1.1 Models without Covariates

In semiparametric maximum likelihood estimation, nonparametric estimates of the marginal survivor functions $S_1(t_1) = Pr(T_1 \geq t_1)$ and $S_2(t_2) = Pr(T_2 \geq t_2)$ and the estimate of α

in the specified parametric copula family $C_\alpha(S_1(t_1), S_2(t_2)) = Pr(T_1 \geq t_1, T_2 \geq t_2)$ are found simultaneously. We maximize (1.6) with respect to S_1, S_2 and α by assuming that the estimates of S_1 and S_2 have jumps only at observed (i.e., uncensored) times t_1 and t_2, respectively. This is a standard assumption in parametric and semiparametric maximum likelihood. It is convenient to use a discrete hazard parametrization to do this, and we define

$$\lambda^*_{11} = 1 - S_1(t^*_{1(1)}),\ \lambda^*_{1l} = [S_1(t^*_{1(l-1)}) - S_1(t^*_{1(l)})]/S_1(t^*_{1(l-1)})\ for\ l = 2, ..., k_1$$

and

$$\lambda^*_{21} = 1 - S_2(t^*_{2(1)}),\ \lambda^*_{2l} = [S_2(t^*_{2(l-1)}) - S_2(t^*_{2(l)})]/S_2(t^*_{2(l-1)})\ for\ l = 2, ..., k_2$$

where $t^*_{1(1)} < t^*_{1(2)} < ... < t^*_{1(k_1)}$ and $t^*_{2(1)} < t^*_{2(2)} < ... < t^*_{2(k_2)}$ are the distinct observed t_{1i}'s with $\delta_{1i} = 1$ and the distinct observed t_{2i}'s with $\delta_{2i} = 1$, respectively, for $i = 1, ..., n$. The likelihood (1.6) can be reexpressed by defining $C^{(1)}_\alpha(u_1, u_2) = \partial C_\alpha(u_1, u_2)/\partial u_1$, $C^{(2)}_\alpha(u_1, u_2) = \partial C_\alpha(u_1, u_2)/\partial u_2$ and $C^{(1,2)}_\alpha(u_1, u_2) = \partial^2 C_\alpha(u_1, u_2)/\partial u_1 \partial u_2$, giving

$$\begin{aligned} L(\lambda^*_1, \lambda^*_2, \alpha) = \prod_{i=1}^{n} & \left[\lambda_{l_1(t_{1i})} S_1(t_{1i}) \lambda_{l_2(t_{2i})} S_2(t_{2i}) C^{(1,2)}_\alpha(S_1(t_{1i}), S_2(t_{2i}))\right]^{\delta_{1i}\delta_{2i}} \\ & \times \left[\lambda_{l_1(t_{1i})} S_1(t_{1i}) C^{(1)}_\alpha(S_1(t_{1i}), S_2(t^+_{2i}))\right]^{\delta_{1i}(1-\delta_{2i})} \\ & \times \left[\lambda_{l_2(t_{2i})} S_2(t_{2i}) C^{(2)}_\alpha(S_1(t^+_{1i}), S_2(t_{2i}))\right]^{(1-\delta_{1i})\delta_{2i}} \\ & \times \left[C_\alpha(S_1(t^+_{1i}), S_2(t^+_{2i}))\right]^{(1-\delta_{1i})(1-\delta_{2i})} \end{aligned} \tag{3.1}$$

where

$$S_1(t_{1i}) = \prod_{l:t^*_{1(l)} < t_{1i}} (1 - \lambda^*_{1l}),\ S_2(t_{2i}) = \prod_{l:t^*_{2(l)} < t_{2i}} (1 - \lambda^*_{2l}),$$

$$S_1(t^+_{1i}) = \prod_{l:t^*_{1(l)} \leq t_{1i}} (1 - \lambda^*_{1l}),\ S_2(t^+_{2i}) = \prod_{l:t^*_{2(l)} \leq t_{2i}} (1 - \lambda^*_{2l}),$$

and where for cases with $\delta_{1i} = 1$, $\lambda_{l_1(t_{1i})}$ is the corresponding λ^*_{1l} where $l_1(t_{1i}) = l : t_{1i} = t^*_{1(l)}$ and for cases with $\delta_{2i} = 1$, $\lambda_{l_2(t_{2i})}$ is the corresponding λ^*_{2l} where $l_2(t_{2i}) = l : t_{2i} = t^*_{2(l)}$. The estimates of S_1, S_2 (i.e. λ^*_1, λ^*_2) and α can be obtained by maximizing the logarithm of (3.1) with general purpose optimizers. Software may run into problems if n is too large. However, for example, $n = 1500$ (without censoring) seem to be feasible to obtain the estimates with the R function nlm.

One can alternatively use an approximate likelihood based on differencing. In this case

the estimates are obtained by maximizing the logarithm of

$$
\begin{aligned}
L(\lambda_1^*, \lambda_2^*, \alpha) = \prod_{i=1}^{n} [& C_\alpha(S_1(t_{1i}), S_2(t_{2i})) - C_\alpha(S_1(t_{1i}), S_2(t_{2i}^+)) \\
& - C_\alpha(S_1(t_{1i}^+), S_2(t_{2i})) + C_\alpha(S_1(t_{1i}^+), S_2(t_{2i}^+))]^{\delta_{1i}\delta_{2i}} \\
& \times \left[C_\alpha(S_1(t_{1i}), S_2(t_{2i}^+)) - C_\alpha(S_1(t_{1i}^+), S_2(t_{2i}^+))\right]^{\delta_{1i}(1-\delta_{2i})} \\
& \times \left[C_\alpha(S_1(t_{1i}^+), S_2(t_{2i})) - C_\alpha(S_1(t_{1i}^+), S_2(t_{2i}^+))\right]^{(1-\delta_{1i})\delta_{2i}} \\
& \times \left[C_\alpha(S_1(t_{1i}^+), S_2(t_{2i}^+))\right]^{(1-\delta_{1i})(1-\delta_{2i})}.
\end{aligned} \tag{3.2}
$$

Although (3.2) approximates the continuous time likelihood (3.1), in fact times are measured discretely in practice and hence, the approximate likelihood is closer to reality.

A nonparametric bootstrap procedure can be used to estimate the variances of estimates of α, $S_1(t_1)$ and $S_2(t_2)$; however, its validity needs to be checked.

Li et al. (2008) developed a semiparametric maximum likelihood estimation procedure based on a normal transformation model. In this model, hazard rate models are transformed to a standard normal model and a joint normal distribution is assumed for a bivariate vector of transformed variates. Similar to our semiparametric maximum likelihood estimation method, they maximize the likelihood function with respect to cumulative hazard functions and correlation parameter in the bivariate normal distribution by assuming that cumulative hazard functions have jumps only at distinct observed failure times. The estimators of the correlation parameter and marginal survivals are shown to be consistent, asymptotically normally distributed and semiparametric efficient under the semiparametric normal transformation model. However, the development of asymptotic theory for the maximum likelihood estimates $\hat{S}_1$, $\hat{S}_2$ and $\hat{\alpha}$ appears difficult in general. Hence, in Section 3.3 we conduct simulation studies based on the approaches done and in the next section.

3.1.2 Models with Proportional Hazards Margins

Copula models with semiparametric Cox models for the marginal distributions have been considered by Glidden and Self (1999), who discuss approximate generalized maximum likelihood estimation for the Clayton copula. Pipper and Martinussen (2003) and Martinussen and Pipper (2005) consider related frailty models with different marginal distributions; they develop approximations to maximum likelihood. Semiparametric maximum likelihood estimation as described in Section 3.1.1 can in fact also be used to fit copula models with proportional hazards margins for data with covariate vectors $\{(x_{1i}, x_{2i}), i = 1, ..., n\}$. In this case, the marginal survivor function of T_j is assumed to be of the form

$$
S_j(t_{ji}) = S_{0j}(t_{ji})^{\exp(\beta' x_{ji})} \tag{3.3}
$$

where the baseline survivor function S_{0j} is arbitrary for $j = 1, 2$ and $i = 1, ..., n$. In some applications including the Diabetic Retinopathy Study of Section 2.3.1 or 3.4.1, the baseline survivor functions are the same (i.e., $S_{01}(t) = S_{02}(t) = S_0(t)$). In any case, (1.6) is maximized with respect to S_{01}, S_{02}, β and α by assuming that the estimates of S_{01} and S_{02} have jumps only at observed times t_1 and t_2, respectively. The estimates can be obtained by maximizing (3.2) where $S_j(t_{ji})$ is as in (3.3), $S_{0j}(t_{ji}) = \prod_{l:t^*_{j(l)}<t_{ji}}(1-\lambda^*_{jl})$, $S_{0j}(t^+_{ji}) = \prod_{l:t^*_{j(l)}\le t_{ji}}(1-\lambda^*_{jl})$, $\lambda^*_{j1} = 1 - S_{0j}(t^*_{j(1)})$, $\lambda^*_{jl} = [S_{0j}(t^*_{j(l-1)}) - S_{0j}(t^*_{j(l)})]/S_{0j}(t^*_{j(l-1)})$ for $l = 2, ..., k_j$ and $j = 1, 2$. We perform this here by using general optimization functions; in particular, in the examples below we used the R function nlm.

It is easy to extend two-stage semiparametric estimation to models with proportional hazards margins (Glidden, 2000). First, a semiparametric proportional hazards model (3.3) is fitted to the marginal distributions with a working independence assumption, giving $\hat{S}_{0j}(t_j)$, $\hat{\beta}$ and, hence, $\hat{S}_j(t_j) = \hat{S}_{0j}(t_j)^{\exp(\hat{\beta}' x_j)}$ for $j = 1, 2$. In the second stage, the vector of dependence parameters α is estimated by maximizing the pseudolikelihood function (1.52). Under some regularity conditions, Glidden (2000) showed that the two-stage estimator of the dependence parameter in the Clayton model with proportional hazards margins is consistent and asymptotically normally distributed.

Two-stage semiparametric estimation is obviously a lot simpler computationally than semiparametric maximum likelihood estimation. Properties of the two estimation methods are investigated in Section 3.3.2 by a simulation study.

3.2 Likelihood Ratio and Pseudolikelihood Ratio Statistics for Goodness of Fit

3.2.1 Models without Covariates

We can use the procedures outlined in Section 3.1 to provide goodness of fit tests for a copula model without making parametric assumptions for the marginal distributions. As before, we use an expanded family of copulas. After fitting the expanded copula model and the proposed copula model under $H_0 : \alpha_2 = \alpha_{20}$ by maximizing $L(\lambda^*_1, \lambda^*_2, \alpha_1, \alpha_2)$ and $L(\lambda^*_1, \lambda^*_2, \alpha_1, \alpha_{20})$ in (3.2) and obtaining the semiparametric maximum likelihood estimates $(\hat{\lambda}^*_1, \hat{\lambda}^*_2, \hat{\alpha}_1, \hat{\alpha}_2)$ and $(\hat{\lambda}^*_1(\alpha_{20}), \hat{\lambda}^*_2(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}))$, respectively, the likelihood ratio test statistic

$$\Lambda_{s1}(\alpha_{20}) = 2\log L(\hat{\lambda}^*_1, \hat{\lambda}^*_2, \hat{\alpha}_1, \hat{\alpha}_2) - 2\log L(\hat{\lambda}^*_1(\alpha_{20}), \hat{\lambda}^*_2(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}), \alpha_{20}) \tag{3.4}$$

can be used to test the null hypothesis. We discuss how to obtain p-values by simulation below.

We can alternatively use a two-stage procedure leading to a pseudolikelihood ratio statistic. Suppose the proposed and the expanded families are fitted by the semiparametric two-stage estimation as described in Section 1.4.1. In other words, in the first stage, the marginal survivor functions are estimated by Kaplan-Meier estimates $\hat{S}_1(t_1)$ and $\hat{S}_2(t_2)$. In the second stage, the vector of dependence parameters $\alpha = (\alpha_1, \alpha_2)$ for the expanded copula family and $\alpha_1(\alpha_{20})$ for the proposed copula family under $H_0 : \alpha_2 = \alpha_{20}$ are estimated by maximizing the pseudolikelihood functions $L_s(\alpha)$ in (1.52) and $L_s(\alpha_1, \alpha_{20})$, respectively, as in (1.53) where C_α is the expanded copula model. Let $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2)$ and $\tilde{\alpha}_1(\alpha_{20})$ be the semiparametric estimates of α and $\alpha_1(\alpha_{20})$, respectively. For testing the null hypothesis, the pseudolikelihood ratio statistic

$$\Lambda_{s2}(\alpha_{20}) = 2 \log L_s(\tilde{\alpha}_1, \tilde{\alpha}_2) - 2 \log L_s(\tilde{\alpha}_1(\alpha_{20}), \alpha_{20}) \tag{3.5}$$

is used. The two-stage semiparametric estimation approach has been shown to give regular asymptotic results for $\tilde{\alpha}$ by Shih and Louis (1995), when α does not lie on a boundary. However, in many settings the parameter value α_{20} lies on a boundary, and in any case the likelihood ratio statistic has a limiting distribution that depends on parameter values.

A semiparametric bootstrap procedure is consequently used to estimate the p-value for testing the null model when the test statistic (3.5) is used. We also apply it with (3.4), since asymptotic theory for the semiparametric likelihood ratio statistic has not yet been established. The steps of the procedure are as follows:

1. Generate data $\{(T_{1i}^*, T_{2i}^*), i = 1, ..., n\}$ from the estimated null model. If semiparametric maximum likelihood was used for fitting the models, we generate $\{(U_{1i}^*, U_{2i}^*), i = 1, ..., n\}$ from $C_{\hat{\alpha}_1(\alpha_{20}),\alpha_{20}}(u_1, u_2)$ and obtain $T_{1i}^* = \hat{S}_1^{-1}(U_{1i}^*)$ and $T_{2i}^* = \hat{S}_2^{-1}(U_{2i}^*)$ where $\hat{S}_1(t_1) = \prod_{l:t_{1(l)}^* < t_1}(1 - \hat{\lambda}_{1l}^*(\alpha_{20}))$ and $\hat{S}_2(t_2) = \prod_{l:t_{2(l)}^* < t_2}(1 - \hat{\lambda}_{2l}^*(\alpha_{20}))$. If the two-stage semiparametric estimation is used, generate $\{(U_{1i}^*, U_{2i}^*), i = 1, ..., n\}$ from $C_{\tilde{\alpha}_1(\alpha_{20}),\alpha_{20}}(u_1, u_2)$ and obtain $T_{1i}^* = \hat{S}_1^{-1}(U_{1i}^*)$ and $T_{2i}^* = \hat{S}_2^{-1}(U_{2i}^*)$ where $\hat{S}_1$ and $\hat{S}_2$ are the Kaplan-Meier estimates of the marginal survivor functions. Note that with probability one, T_{1i}^* and T_{2i}^* are uniquely determined.

2. Generate censoring times $\{(C_{1i}^*, C_{2i}^*), i = 1, ..., n\}$ from an estimate of the distribution of (C_1, C_2) according to the properties of censoring in the given data set.

3. Compute $t_{1i}^* = \min(T_{1i}^*, C_{1i}^*)$, $t_{2i}^* = \min(T_{2i}^*, C_{2i}^*)$, $\delta_{1i}^* = I[T_{1i}^* = t_{1i}^*]$ and $\delta_{2i}^* = I[T_{2i}^* = t_{2i}^*]$ for $i = 1, ..., n$.

4. If (3.4) is being used to test the null hypothesis, obtain the semiparametric maximum likelihood estimates $(\hat{\lambda}_1^*, \hat{\lambda}_2^*, \hat{\alpha}_1^*, \hat{\alpha}_2^*)$ and $(\hat{\lambda}_1^*(\alpha_{20}), \hat{\lambda}_2^*(\alpha_{20}), \hat{\alpha}_1^*(\alpha_{20}))$ under the expanded and null models, respectively, by maximizing the likelihood function (3.2) for the bootstrap sample $\{(t_{1i}^*, \delta_{1i}^*, t_{2i}^*, \delta_{2i}^*), i = 1, ..., n\}$.

If (3.5) is used to test the null hypothesis, for the bootstrap sample, in the first stage obtain the Kaplan-Meier estimates $\hat{S}_1^*$ and $\hat{S}_2^*$ from $\{(t_{1i}^*, \delta_{1i}^*), i = 1, ..., n\}$ and $\{(t_{2i}^*, \delta_{2i}^*), i = 1, ..., n\}$, respectively. In the second stage, find the estimates $\tilde{\alpha}^* = (\tilde{\alpha}_1^*, \tilde{\alpha}_2^*)$ and $\tilde{\alpha}_1^*(\alpha_{20})$ for the expanded and proposed copula models by maximizing the pseudolikelihood function (1.52), respectively.

5. Calculate the value Λ_s^* of the test statistic (3.4) or (3.5) based on the bootstrap sample.

6. Steps 1 to 5 are repeated B times and the p-value is calculated as the proportion of times that $\Lambda_s^* \geq \Lambda_s^{obs}$, where Λ_s^{obs} is the observed value of $\Lambda_{s1}(\alpha_{20})$ or $\Lambda_{s2}(\alpha_{20})$.

As mentioned in Section 1.5.3, Chen and Huang (2007) used another semiparametric bootstrap procedure. Both bootstrap procedures appear valid, though no theoretical development for either is available. Limited simulation results for our procedure suggest that it provides satisfactory p-values for samples of size 100 in the settings of Section 3.3.

3.2.2 Models with Proportional Hazards Margins

If the data has covariates and the expanded and proposed copula models with proportional hazards margins are fitted by semiparametric maximum likelihood estimation, then the likelihood ratio test statistic

$$\Lambda_{s1}(\alpha_{20}) = 2 \log L(\hat{\beta}, \hat{\lambda}_1^*, \hat{\lambda}_2^*, \hat{\alpha}_1, \hat{\alpha}_2) - 2 \log L(\hat{\beta}(\alpha_{20}), \hat{\lambda}_1^*(\alpha_{20}), \hat{\lambda}_2^*(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}), \alpha_{20}) \quad (3.6)$$

can be used to test the null hypothesis $H_0 : \alpha_2 = \alpha_{20}$. If the models are fitted by two-stage semiparametric estimation as described in Section 3.1.2, the pseudolikelihood ratio statistic is evaluated as in (3.5).

To obtain p-values we propose a semiparametric bootstrap procedure, the only change from the one in Section 3.2.1 being when generating data from the estimated null model in the first step. If semiparametric maximum likelihood was used for fitting the models, after generating $\{(U_{1i}^*, U_{2i}^*), i = 1, ..., n\}$ from $C_{\hat{\alpha}_1(\alpha_{20}),\alpha_{20}}(u_1, u_2)$, we obtain $T_{1i}^* = \hat{S}_{01}^{-1}(U_{1i}^{* \exp(-\hat{\beta}(\alpha_{20})' x_{1i})})$ and $T_{2i}^* = \hat{S}_{02}^{-1}(U_{2i}^{* \exp(-\hat{\beta}(\alpha_{20})' x_{2i})})$ where $\hat{S}_{01}(t_1) = \prod_{l:t_{1(l)}^* < t_1}(1 - \hat{\lambda}_{1l}^*(\alpha_{20}))$ and $\hat{S}_{02}(t_2) = \prod_{l:t_{2(l)}^* < t_2}(1 - \hat{\lambda}_{2l}^*(\alpha_{20}))$. If two-stage semiparametric estimation was used, after generating $\{(U_{1i}^*, U_{2i}^*), i = 1, ..., n\}$ from $C_{\tilde{\alpha}_1(\alpha_{20}),\alpha_{20}}(u_1, u_2)$, obtain $T_{1i}^* = \hat{S}_{01}^{-1}(U_{1i}^{* \exp(-\hat{\beta}' x_{1i})})$ and $T_{2i}^* = \hat{S}_{02}^{-1}(U_{2i}^{* \exp(-\hat{\beta}' x_{2i})})$ where $\hat{S}_{01}$ and $\hat{S}_{02}$ are the estimates of baseline survivor functions S_{01} and S_{02} under the working independence assumption.

n	% Censored	τ	True copula	Proportion
100	0	0.4	$\theta = 1$, $\phi = 1.333$	0.031
100	0	0.8	$\theta = 1$, $\phi = 8$	0.050
100	30	0.4	$\theta = 1$, $\phi = 1.333$	0.054
100	30	0.8	$\theta = 1$, $\phi = 8$	0.041
100	50	0.4	$\theta = 1$, $\phi = 1.333$	0.045
100	50	0.8	$\theta = 1$, $\phi = 8$	0.038
100	0	0.4	$\phi = 0$, $\theta = 1.667$	0.967
100	0	0.8	$\phi = 0$, $\theta = 5$	1
100	30	0.4	$\phi = 0$, $\theta = 1.667$	0.912
100	30	0.8	$\phi = 0$, $\theta = 5$	0.996
100	50	0.4	$\phi = 0$, $\theta = 1.667$	0.782
100	50	0.8	$\phi = 0$, $\theta = 5$	0.943

Table 3.1: Proportion of rejections of $H_0 : \theta = 1$ (i.e., Clayton family), under models (1.40), for the pseudolikelihood ratio statistic.

3.3 Simulation Study

3.3.1 Performance of Semiparametric Pseudolikelihood Ratio Statistic in Testing the Clayton Copula

The full semiparametric maximum likelihood estimation requires considerable computer time with larger sample sizes, and given the absence of large sample approximations, we choose for now to consider only the two-stage approach for tests of fit, since we need to get p-values by simulation. A simulation study was carried out with the same design as in Section 2.2 (i.e., with the same 1000 random bivariate failure time samples of size 100 generated from members of the true copula family (1.40)) to assess the performance of the semiparametric pseudolikelihood ratio test statistic Λ_{s2} in (3.5). As earlier, critical values for Λ_{s2} were estimated from 10000 independent samples for each scenario. The empirical type I errors and power values corresponding to those in Table 2.1 are given in Table 3.1. We observe that empirical type I errors are generally close to 0.05 except for the case where there is no censoring and Kendall's tau is 0.4. The semiparametric pseudolikelihood ratio test is almost as powerful as the parametric likelihood ratio and pseudolikelihood ratio tests (based on correct marginal specifications) shown in Table 2.1 while achieving robustness to the form of the marginal distributions.

The power of the semiparametric pseudolikelihood ratio test was also investigated when the true copula model is not a member of the expanded copula family. We performed the

n	% Censored	τ	Proportion
100	0	0.4	0.764
100	0	0.8	0.987
100	30	0.4	0.458
100	30	0.8	0.664
100	50	0.4	0.288
100	50	0.8	0.362

Table 3.2: Proportion of rejections of $H_0 : \theta = 1$ (i.e., Clayton family) for a test based on (1.40) but with (1.34) the true copula.

same simulation study as above but assuming the true copula model is the Frank, with two degrees of association represented by Kendall's tau values of $\tau = 0.4$ and 0.8, which correspond to $\nu = 4.16$ and $\nu = 18.2$ in (1.34), respectively. The empirical powers of the test are given in Table 3.2 when the true copula model is the Frank (1.34), the null model is the Clayton (1.28) and the misspecified expanded model is (1.40). When we compare the results with those given in Table 2.2, we observe that the empirical power values of the semiparametric pseudolikelihood ratio test are a little lower than those of the parametric likelihood and pseudolikelihood ratio tests, but the latter rely on correct marginal specifications. The comments from Section 2.2 concerning the difficulties of detecting departures from a bivariate copula model with heavily censored samples of size 100 naturally apply here also.

3.3.2 Performance of Semiparametric Maximum Likelihood and Two-Stage Semiparametric Estimators

Semiparametric maximum likelihood is less attractive for testing fit because, as noted above, simulation is needed to obtain p-values and the procedure is computationally demanding. However, it is of interest to compare the efficiency of estimation of a copula dependence parameter α with the one- and two-stage methods. The performance of the semiparametric maximum likelihood estimator of the association parameter is compared here with that of the two-stage semiparametric estimator of it. We did a simulation study similar to that in Shih and Louis (1995). We consider Clayton, Gumbel-Hougaard and Frank copulas, with two degrees of association represented by Kendall's tau values of 0.4 and 0.7 for each. The marginal distributions of the failure times are considered as Exponential with a unit scale parameter. We consider both uncensored and censored samples. For the censored case, the bivariate censoring times $C_{1i},\ C_{2i}$ were generated independently from Uniform distribution over $(0, 3.2)$ so that the probability of censoring in each co-

ordinate is 30%. For each case, we generated 500 simulated samples with $n = 50$ and 100.

In Table 3.3, empirical biases and standard deviations of the semiparametric maximum likelihood and the two-stage semiparametric estimates of a reparametrized version of the parameter in the Clayton (1.28), Gumbel-Hougaard (1.31) and Frank copula (1.34) models are given. We estimated $\gamma = \log \phi$ for the Clayton copula, $\gamma = \log(\theta - 1)$ for the Gumbel-Hougaard copula and $\gamma = \log \nu$ for the Frank copula. For both uncensored and censored cases, the size of the bias relative to the standard deviation of the semiparametric maximum likelihood estimator is lower than that of the two-stage semiparametric estimator when there is moderate association (i.e., $\tau = 0.4$) in all of the copulas. When Kendall's tau is 0.7, the size of the bias relative to the standard deviation of the semiparametric maximum likelihood estimator of the Gumbel-Hougaard copula parameter is again generally lower, however, that of the Clayton and Frank parameters are higher. It is observed that the two-stage semiparametric estimator of the copula parameter is about as good as the semiparametric maximum likelihood estimator of it.

We also checked on the asymptotic normality of semiparametric maximum likelihood and two-stage semiparametric estimators of γ by applying Anderson-Darling test statistic to the 500 estimates $\hat{\gamma}$ obtained for each senario. For both uncensored and censored cases, when the true model is Frank or when it is Clayton or Gumbel-Hougaard with Kendall's tau 0.7, it is observed that the approximate normality is reasonable even when sample size is 50. When the true model is Clayton or Gumbel-Hougaard with moderate association ($\tau = 0.4$), a sample size of 100 or more is needed for the estimators to be close to normal. We remark, however, that we have examined the $\hat{\gamma}$'s; if we had a variance estimate we might find that standardized Wald statistics $(\hat{\gamma} - \gamma)/\sqrt{\hat{V}ar(\hat{\gamma})}$ were closer to normally distributed.

3.3.3 Asymptotic Distributions of Semiparametric Likelihood Ratio and Pseudolikelihood Ratio Statistics

Work on semiparametric maximum likelihood by Murphy and van der Vaart (2000) suggests that likelihood ratio statistics and pseudolikelihood ratio statistics about copula parameters might have chi-squared asymptotics similar to those for fully parametric models, for cases where the marginal distributions are non- or semi-parametrically specified. In this section, we examine properties of semiparametric likelihood ratio and pseudolikelihood ratio statistics by examining the likelihood ratio statistic values obtained from continuation of the simulation study given in Section 3.3.2. We consider the null hypothesis $H_0 : \gamma = \gamma_0$ where γ_0 is the true parameter value of the corresponding copula model by using semiparametric likelihood and pseudolikelihood ratio statistics $\Lambda_{s1}(\gamma_0)$ and $\Lambda_{s2}(\gamma_0)$.

				No censoring		30% censoring	
Model	τ	True γ		$n = 50$	$n = 100$	$n = 50$	$n = 100$
Clayton	0.4	0.288	(a)	0.007 (0.363)	0.003 (0.245)	0.033 (0.472)	0.008 (0.302)
			(b)	0.024 (0.337)	0.009 (0.232)	-0.037 (0.629)	-0.022 (0.301)
	0.7	1.540	(a)	0.089 (0.268)	0.052 (0.175)	0.131 (0.342)	0.063 (0.217)
			(b)	-0.029 (0.250)	-0.023 (0.169)	-0.082 (0.302)	-0.047 (0.211)
Gumbel-	0.4	-0.405	(a)	0.017 (0.382)	0.011 (0.272)	0.048 (0.446)	0.024 (0.300)
Hougaard			(b)	-0.067 (0.389)	-0.043 (0.279)	-0.086 (0.455)	-0.057 (0.307)
	0.7	0.847	(a)	0.045 (0.241)	0.034 (0.154)	0.100 (0.313)	0.084 (0.194)
			(b)	-0.102 (0.231)	-0.058 (0.152)	-0.148 (0.290)	-0.071 (0.190)
Frank	0.4	1.426	(a)	0.009 (0.271)	0.009 (0.187)	0.013 (0.313)	0.010 (0.220)
			(b)	-0.037 (0.281)	-0.013 (0.190)	-0.050 (0.324)	-0.022 (0.224)
	0.7	2.434	(a)	0.056 (0.175)	0.036 (0.116)	0.091 (0.228)	0.044 (0.149)
			(b)	-0.049 (0.175)	-0.017 (0.114)	-0.061 (0.204)	-0.032 (0.140)

Table 3.3: Empirical biases and standard deviations (given in parenthesis) of (a) semiparametric maximum likelihood estimate and (b) two-stage semiparametric estimate of copula parameter computed from 500 samples.

To check whether the semiparametric likelihood ratio statistic has an asymptotic chi-squared distribution with 1 degree of freedom, we used the Kolmogorov-Simirnov test statistic. For both uncensored and censored cases, when the true copula model is Frank with any Kendall's tau and Clayton with Kendall's tau 0.4 and Gumbel-Hougaard with Kendall's tau 0.7, the chi-squared distribution assumption seems to be appropriate.

In addition, $p \times 100 = 90$, 95 and 99th quantiles $Q(p)$ of a chi-squared distribution with degree of freedom 1 and empirical values of $Pr(\Lambda_{s1} > Q(p)) = 1 - p$ and $Q(p)$ computed from 500 samples are given in Table 3.4. $'*'$ indicates that the corresponding empirical value of $1 - p$ does not fall in a 95% confidence interval for $1 - p$. The results in Table 3.4 generally coincide with the ones obtained from the Kolmogorov-Simirnov test.

In Figures 3.1, 3.2 and 3.3, the quantile-quantile plots of semiparametric likelihood and pseudolikelihood ratio statistics are given when the true copula models are Clayton, Gumbel-Hougaard and Frank, respectively, for uncensored samples with size $n = 100$. They show that an adjustment for the semiparametric likelihood and pseudolikelihood ratio statistics is generally necessary to obtain that the distribution of the statistics is approximated by a chi-squared distribution with 1 degree of freedom. However, since it is hard to find the correction term for each copula model seperately, especially under censoring, and since correction terms may depend on the unknown parameters and a parameter can be on the boundary as illustrated in Section 2.1.1, it is suggested to estimate p-values by using the bootstrap procedure described in Section 3.2.

3.4 Applications

Diabetic Rethinopathy Study data and insurance data were analyzed in Section 2.3 parametrically, and are analyzed semiparametrically in this section. It can be seen from the following that semiparametric estimates of the dependence parameters are very close to their fully parametric estimates and the same conclusions are reached as before for both of the data sets.

3.4.1 Diabetic Rethinopathy Study Data

The expanded and proposed copula models were fitted here by semiparametric maximum likelihood estimation when the marginal distributions are modeled by a Cox semiparametric proportional hazards model (3.3) where the two eyes of an individual have the same baseline survivor function (i.e., $S_{01} = S_{02} = S_0$), that is by (2.14) where the baseline hazard function is arbitrary. In Table 3.5, the semiparametric maximum likelihood estimates of the parameters and maximized log-likelihood values are shown. For the Clayton model,

			True		Clayton		Gumbel-Hougaard		Frank	
n	% Cens.	τ	p	$Q(p)$	$1-\hat{p}$	$\hat{Q}(p)$	$1-\hat{p}$	$\hat{Q}(p)$	$1-\hat{p}$	$\hat{Q}(p)$
50	0	0.4	0.90	2.706	0.120	2.950	0.126	3.222	0.122	3.122
			0.95	3.841	0.074*	4.434	0.062	4.548	0.066	4.302
			0.99	6.635	0.018	6.832	0.024*	10.077	0.014	7.761
100	0	0.4	0.90	2.706	0.110	2.914	0.144*	3.342	0.124	3.014
			0.95	3.841	0.064	4.251	0.074*	4.989	0.056	3.959
			0.99	6.635	0.014	8.515	0.026*	9.061	0.018	7.348
50	30	0.4	0.90	2.706	0.124	3.022	0.130*	3.240	0.118	2.870
			0.95	3.841	0.076*	5.182	0.076*	4.880	0.062	4.125
			0.99	6.635	0.028*	7.462	0.034*	9.787	0.006	6.067
100	30	0.4	0.90	2.706	0.124	2.910	0.138*	3.441	0.122	2.984
			0.95	3.841	0.062	4.313	0.080*	5.118	0.056	4.120
			0.99	6.635	0.016	7.472	0.018	8.367	0.022*	7.447
50	0	0.7	0.90	2.706	0.174*	3.832	0.110	2.766	0.134*	3.252
			0.95	3.841	0.100*	5.439	0.070*	4.246	0.070*	4.417
			0.99	6.635	0.032*	8.745	0.020*	7.657	0.018	7.400
100	0	0.7	0.90	2.706	0.148*	3.476	0.082	2.473	0.136*	3.355
			0.95	3.841	0.078*	4.777	0.044	3.561	0.078*	4.589
			0.99	6.635	0.022*	8.248	0.008	6.257	0.010	6.310
50	30	0.7	0.90	2.706	0.160*	3.456	0.152*	3.604	0.150*	3.550
			0.95	3.841	0.090*	6.194	0.086*	5.662	0.084*	5.146
			0.99	6.635	0.050*	9.861	0.036*	8.957	0.022*	8.640
100	30	0.7	0.90	2.706	0.120	3.116	0.142*	3.335	0.128*	3.105
			0.95	3.841	0.072*	4.462	0.066	4.393	0.072*	4.634
			0.99	6.635	0.022*	9.538	0.028*	8.241	0.018	8.438

Table 3.4: $p \times 100 = 90$, 95 and 99th quantiles $Q(p)$ of a chi-squared distribution with degrees of freedom 1 and empirical values of $Pr(\Lambda_{s1} > Q(p)) = 1 - p$ and $Q(p)$ computed from 500 samples.

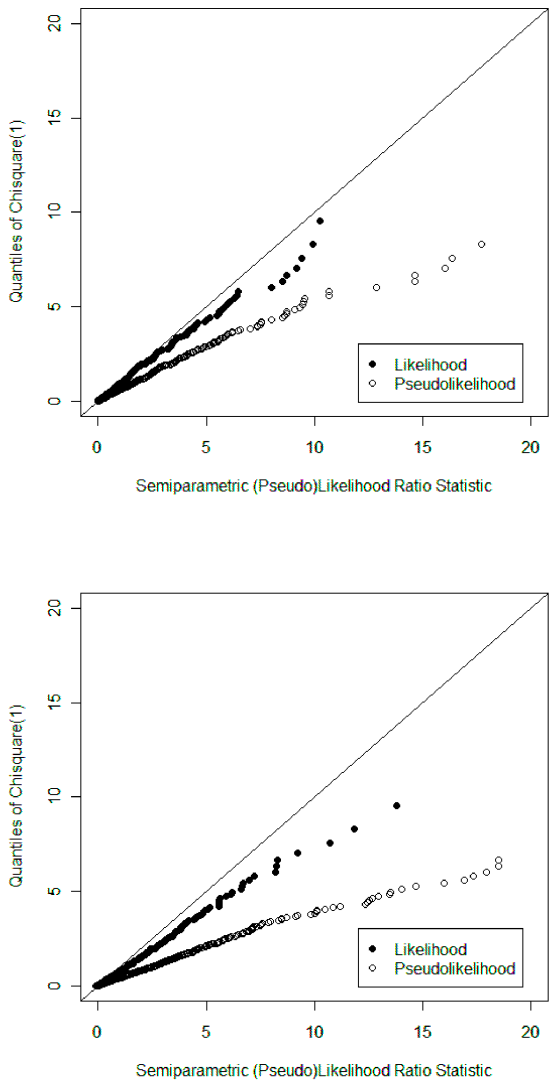
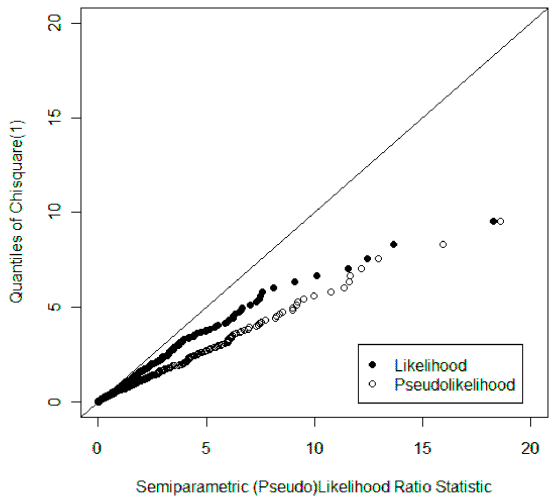


Figure 3.1: Quantile-quantile plots of semiparametric likelihood and pseudolikelihood ratio statistics when the true copula model is Clayton with Kendall's tau 0.4 (top plot) and 0.7 (bottom plot), sample size is $n = 100$ and there is no censoring.



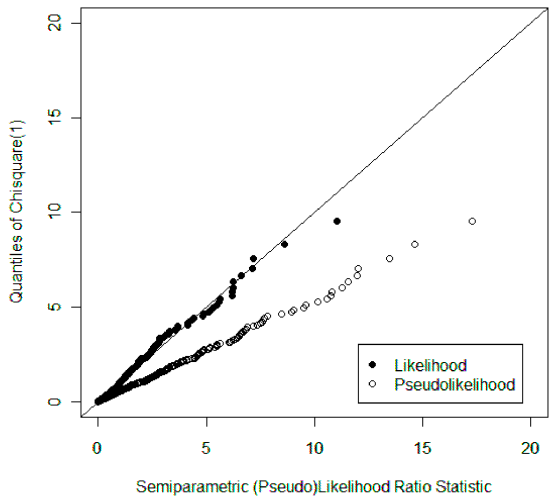


Figure 3.2: Quantile-quantile plots of semiparametric likelihood and pseudolikelihood ratio statistics when the true copula model is Gumbel-Hougaard with Kendall's tau 0.4 (top plot) and 0.7 (bottom plot), sample size is $n = 100$ and there is no censoring.

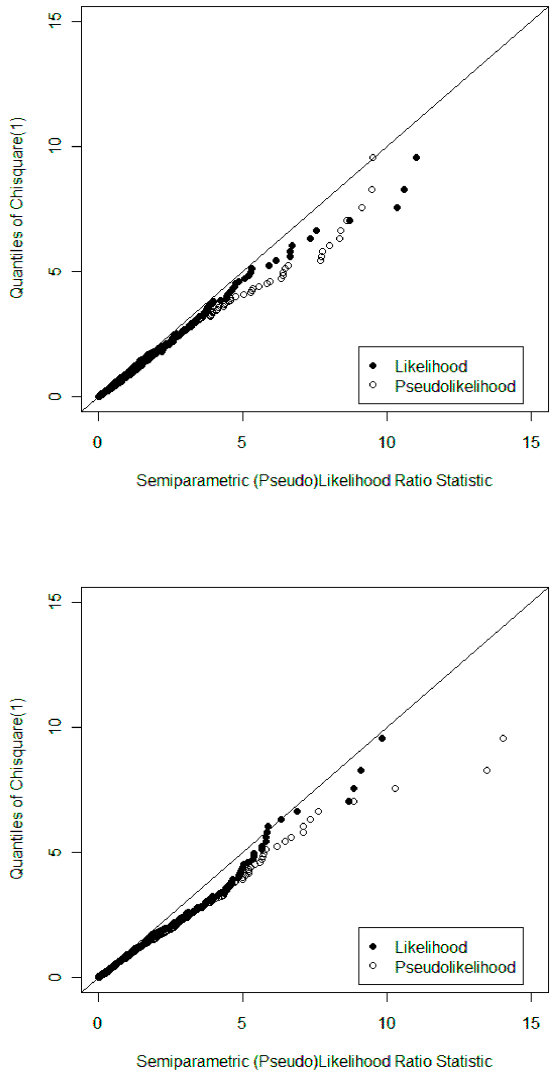


Figure 3.3: Quantile-quantile plots of semiparametric likelihood and pseudolikelihood ratio statistics when the true copula model is Frank with Kendall's tau 0.4 (top plot) and 0.7 (bottom plot), sample size is $n = 100$ and there is no censoring.

Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Two-parameter	-974.387	0.934	1.031	-0.423	0.353	-0.819
Clayton	-974.427	1.068	1	-0.421	0.354	-0.816
Gumbel-Hougaard	-976.033	0	1.238	-0.433	0.352	-0.812

Table 3.5: Semiparametric maximum likelihood estimation results for DRS data.

Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$
Working independence		0	1	-0.425	0.341	-0.846
Two-parameter	-101.350	0.791	1.075			
Clayton	-101.610	1.107	1			
Gumbel-Hougaard	-102.761	0	1.239			

Table 3.6: Two-stage semiparametric estimation results for DRS data.

our estimates are very close to those of Glidden and Self (1999) who used an approximate maximum likelihood approach. For testing the Clayton and the Gumbel-Hougaard copula models, likelihood ratio statistics were found as $\Lambda_{s1}(1) = 0.080$ and $\Lambda_{s1}(0) = 3.290$, respectively. When the semiparametric bootstrap procedure described in Section 3.2.2 is applied, the p-values are estimated as 0.273 for the Clayton model and 0.049 for the Gumbel-Hougaard model. There is some mild evidence against the Gumbel-Hougaard model. In this case we have not been constrained by the assumption of Weibull marginal distributions, but proportional hazards is still assumed.

We also fitted the copula families with two-stage semiparametric estimation. The two-stage semiparametric estimates of the dependence parameters and the maximized log-likelihood values are given in Table 3.6. Pseudolikelihood ratio statistics were found as $\Lambda_{s2}(1) = 0.521$ and $\Lambda_{s2}(0) = 2.822$ for testing the Clayton and Gumbel-Hougaard copula models, respectively. We carried out the semiparametric bootstrap procedure to obtain the p-values for testing the proposed models and they are estimated as 0.18 for the Clayton model and 0.102 for the Gumbel-Hougaard model. The same conclusion that the Clayton model fits slightly better is reached as when parametric likelihood or pseudolikelihood ratio tests are conducted in Section 2.3.1.

Note that to estimate the variances of estimates of the dependence parameters or to get confidence intervals for them, a nonparametric bootstrap procedure can be used.

3.4.2 Insurance Data

Two-stage semiparametric estimation (see Table 3.7) yields pseudolikelihood ratio statistics of 221.6 (Clayton model) and virtually zero (Gumbel-Hougaard model), very close to those for the parametric case. Note that in this case, full semiparametric maximum likelihood is computationally forbidding because of the size of n.

Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$
Two-parameter	115.481	6×10^{-5}	1.441
Clayton	4.662	0.488	1
Gumbel-Hougaard	115.481	0	1.441

Table 3.7: Two-stage semiparametric estimation results for insurance loss data.

Chapter 4

Estimation and Tests of Fit Based on Sequential Lifetime Data

In many settings one encounters sequences of survival times, observed one after the other. Difficulties in modeling and analyzing sequential lifetime data have become well known (e.g. see Cook and Lawless, 2007, Chapter 4; Lawless and Fong, 1999; Schaubel and Cai, 2004ab). In particular, followup studies in which a sequence of survival times may be observed are of finite duration, with the result that survival times can be censored. However, because the survival times for a given individual are typically not independent and because they are observed sequentially, a form of dependent censoring is induced even when the overall followup time C is independently determined. For example, if the study is of a fixed length C and the first survival time is T_1, then a censoring time $C_2 = C - T_1$ is induced for the second survival time, T_2. If T_1 and T_2 are related then so are T_2 and C_2. Thus independent censoring does not apply to T_2 and simple methods of analysis that focus on the marginal distribution of T_2 and related covariates cannot be applied. Moreover, there is a fundamental identifiability problem in most studies, related to the fact that second or subsequent survival times are observable only if preceding survival times for an individual are uncensored (Lin et al., 1999; Schaubel and Cai, 2004a; Cook and Lawless, 2007, Section 4.4.1). In addition, in some studies, a significant proportion of individuals do not experience the first event or survival time. For example, in the case of cancer relapse and death in the example given in Section 1.1.2, some individuals may not suffer a relapse and, indeed, may be cured of their disease (i.e., $p = F_1(\infty) < 1$). Thus, modeling of the distribution of T_1 and the joint distribution of T_1, T_2 should reflect this feature.

As discussed in Section 1.4.2, a number of authors have studied nonparametric estimation of $F(t_1, t_2)$ in the case in which there are no covariates. In fact, unless T_1 has finite support with $Pr(T_1 \leq C_{\max}) = 1$, where $C_{\max}$ is the largest followup time in a study, the best

that can be estimated nonparametrically are probabilities $F(t_1, t_2)$, where $t_1 + t_2 \leq C_{\max}$. Correspondingly, although $F_1(t_1) = Pr(T_1 \leq t_1)$ is estimable for $0 \leq t_1 \leq C_{\max}$, for T_2 all that is estimable are conditional probabilities $Pr(T_2 \leq t_2 | T_1 \leq t_1)$, where $t_1 + t_2 \leq C_{\max}$. Lin et al. (1999), Schaubel and Cai (2004a) and Cook and Lawless (2007, Section 4.4.1) provide good discussions of this issue.

There is often considerable interest in estimating the marginal distribution for T_2. One way to do this is to adopt a parametric model for the joint distribution of T_1 and T_2. This approach was taken by He and Lawless (2003), who used a copula formulation along with piecewise constant or spline hazard functions for T_1 and T_2, to reduce reliance on strong parametric assumptions. However, the approach is still parametric, and can sometimes have trouble picking up the shapes of the hazard functions. Moreover, ways of checking these or simple parametric models for sequential survival times are currently lacking. Our objective is to propose new semiparametric approaches, in which a copula is used to model association between T_1 and T_2, but the marginal distributions of T_1 and T_2 are left nonparametric. Our models also incorporate the possibility that $F_1(t_1)$ approaches a value $p < 1$ as $t_1 \to \infty$; this allows for the feature discussed in the cancer treatment example in Section 1.1.2, where some individuals are cured and will never experience a relapse.

In the following section, we present our modelling approach and show how the case where $p = F_1(\infty) < 1$ is handled. Section 4.2 develops some semiparametric estimation procedures. Section 4.3 presents a simulation study demonstrating their properties. In sections 4.4 and 4.5 we introduce another modelling approach and a semiparametric estimation technique to fit the model, respectively. Section 4.6 applies the methodology developed to the colon cancer treatment data.

4.1 Copula Models for a Sequence of Survival Times

In some sequential bivariate data such as the colon cancer data described in Section 1.1.2, some individuals never have the first event, which defines T_1. For example, a fraction $1 - p$ of patients might have no chance of disease recurrence. In these settings, a mixture model is appropriate to represent the distribution of the time to first event (Lawless, 2003). Therefore, for $t_1 < \infty$, the distribution function for the time to the first event is

$$F_1(t_1) = Pr(T_1 \leq t_1) = pPr(T_1 \leq t_1 | T_1 < \infty) = pF_0(t_1) \tag{4.1}$$

where $0 < p = Pr(T_1 < \infty) \leq 1$ and $F_0(t_1)$ is a conditional distribution function of T_1 given $T_1 < \infty$ with $F_0(0) = 0$ and $F_0(\infty) = 1$. Note that the case $p = 1$ is also included in this type of model.

Since the second survival time T_2 can only be observed when the first has been observed, that is if $T_1 < \infty$, the distribution function of T_2 is modelled as

$$F_2(t_2) = Pr(T_2 \leq t_2 | T_1 < \infty). \tag{4.2}$$

We consider two approaches to model the bivariate distribution of the successive survival times T_1 and T_2 by using copula models. In the first approach, the bivariate distribution of T_1 and T_2 is modelled as

$$F(t_1, t_2) = Pr(T_1 \leq t_1, T_2 \leq t_2) = C'(F_1(t_1), F_2(t_2)) \tag{4.3}$$

for $t_1 < \infty$ with $F(\infty, \infty) = p$. The properties of the function C' are given in Theorem 4 below.

Assume $F_1(t_1)$ is strictly monotone increasing for $t_1 \geq 0$. The distribution of $F_1(T_1)$ is $Uniform(0, p)$ as it is shown in Theorem 3. The effect of that on the properties of the model (4.3) is presented in Theorem 4.

Theorem 3. *Given $U_1 < p$ ($T_1 < \infty$), $U_1 = F_1(T_1)$ is distributed as $Uniform(0, p)$.*

Proof.

$$Pr(U_1 \leq u_1 | U_1 < p) = \frac{1}{p} Pr(F_1(T_1) \leq u_1) = \frac{1}{p} Pr(T_1 \leq F_1^{-1}(u_1)) = \frac{1}{p} F_1(F_1^{-1}(u_1)) = \frac{u_1}{p}$$

for $0 \leq u_1 \leq p$. □

Theorem 4. *The function $C'(u_1, u_2)$ in (4.3) is defined on $[0, p] \times [0, 1]$ and has the following properties:*

1. *$C'(p, u_2) = pu_2$, $0 \leq u_2 \leq 1$.*
2. *$C'(u_1, 1) = u_1$, $0 \leq u_1 \leq p$.*
3. *$C'(0, u_2) = C'(u_1, 0) = 0$, $0 \leq u_1 \leq p$ and $0 \leq u_2 \leq 1$.*
4. *$C'(v_1, v_2) - C'(v_1, u_2) - C'(u_1, v_2) + C'(u_1, u_2) \geq 0$ whenever $(u_1, u_2) \in [0, p] \times [0, 1]$ and $(v_1, v_2) \in [0, p] \times [0, 1]$ such that $0 \leq u_1 \leq v_1 \leq p$ and $0 \leq u_2 \leq v_2 \leq 1$.*

Proof. 1. $C'(p, u_2) = Pr(F_1(T_1) \leq p, F_2(T_2) \leq u_2) = pPr(F_2(T_2) \leq u_2 | F_1(T_1) \leq p) = pPr(T_2 \leq F_2^{-1}(u_2) | T_1 < \infty) = pF_2(F_2^{-1}(u_2)) = pu_2$.

2. $C'(u_1, 1) = Pr(F_1(T_1) \leq u_1, F_2(T_2) \leq 1) = Pr(F_1(T_1) \leq u_1) = u_1$.

3. $C'(0,u_2) = Pr(F_1(T_1) \leq 0, F_2(T_2) \leq u_2) = 0$ and similarly, $C'(u_1,0) = Pr(F_1(T_1) \leq u_1, F_2(T_2) \leq 0) = 0$.

4. $C'(v_1,v_2) - C'(v_1,u_2) - C'(u_1,v_2) + C'(u_1,u_2) = Pr(F_1(T_1) \leq v_1, F_2(T_2) \leq v_2) - Pr(F_1(T_1) \leq v_1, F_2(T_2) \leq u_2) - Pr(F_1(T_1) \leq u_1, F_2(T_2) \leq v_2) + Pr(F_1(T_1) \leq u_1, F_2(T_2) \leq u_2) \geq 0$.

□

We now consider estimation based on data from n independent individuals. The likelihood function in (1.7) is written in terms of $C'(F_1(t_1), F_2(t_2))$ as follows:

$$L = \prod_{i=1}^{n} \left[\frac{\partial^2 C'(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial C'(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} \times [1 - F_1(t_{1i})]^{1-\delta_{1i}}. \tag{4.4}$$

We discuss estimation based on this below.

It may sometimes be better to represent the bivariate distribution of the times (T_1, T_2) in the semi-survival form

$$Pr(T_1 \leq t_1, T_2 > t_2) = C'(F_1(t_1), S_2(t_2)) \tag{4.5}$$

for $t_1 < \infty$, where $C'(u_1,u_2)$ is a function defined on $0 \leq u_1 \leq p$, $0 \leq u_2 \leq 1$, $F_1(t_1)$ is given in (4.1) and $S_2(t_2) = 1 - F_2(t_2)$ is the survivor function of T_2 given $T_1 < \infty$ with $S_2(0) = 1$, $S_2(\infty) = 0$. The likelihood function in (1.7) and (4.4) is written in terms of $C'(F_1(t_1), S_2(t_2))$ as follows:

$$L = \prod_{i=1}^{n} \left[-\frac{\partial^2 C'(F_1(t_{1i}), S_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial C'(F_1(t_{1i}), S_2(t_{2i}))}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} [1 - F_1(t_{1i})]^{1-\delta_{1i}}. \tag{4.6}$$

Theorem 4 shows C' is not a copula, but we can re-express the models (4.3) and (4.5) using copulas. To do this, for (4.3) we model $Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 < \infty)$ with a standard copula function C as

$$Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 < \infty) = C(F_0(t_1), F_2(t_2)), \tag{4.7}$$

where $F_0(t_1)$ is described in (4.1) and $F_2(t_2)$ is given in (4.2). The likelihood function in (1.7) is written in terms of $C(F_0(t_1), F_2(t_2))$ as follows:

$$L = \prod_{i=1}^{n} p^{\delta_{1i}} \left[\frac{\partial^2 C(F_0(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial F_0(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_0(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} \times [1 - pF_0(t_{1i})]^{1-\delta_{1i}}. \tag{4.8}$$

For the model (4.5) we consider

$$Pr(T_1 \leq t_1, T_2 > t_2 | T_1 < \infty) = C(F_0(t_1), S_2(t_2)), \tag{4.9}$$

and then the likelihood function (4.6) is written as

$$L = \prod_{i=1}^{n} p^{\delta_{1i}} \left[-\frac{\partial^2 C(F_0(t_{1i}), S_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial C(F_0(t_{1i}), S_2(t_{2i}))}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} [1 - pF_0(t_{1i})]^{1-\delta_{1i}}. \tag{4.10}$$

The model in (4.3) can thus be represented in terms of a cdf copula (4.7) as

$$Pr(T_1 \leq t_1, T_2 \leq t_2, T_1 < \infty) = C'(F_1(t_1), F_2(t_2)) = pC(F_0(t_1), F_2(t_2)) \tag{4.11}$$

and the model in (4.5) can be represented in terms of a semi-survival copula (4.9) as

$$Pr(T_1 \leq t_1, T_2 > t_2, T_1 < \infty) = C'(F_1(t_1), S_2(t_2)) = pC(F_0(t_1), S_2(t_2)). \tag{4.12}$$

Many well known copula models allow positive association only between U_1 and U_2. When there is a positive association between T_1 and T_2, the model given in (4.11) is then useful. However, the semi-survival model given in (4.12) is useful if the association between T_1 and T_2 is negative.

In the following section, the most relevant way to describe the association between the times T_1 and T_2 is considered when $F_1(\infty) = p < 1$, and a commonly used association measure Kendall's tau is investigated.

4.1.1 Dependence Measures

The most relevant association measure is based on the association between T_1 and T_2 given $T_1 < \infty$ as we can only observe T_2 if T_1 is observed. Let (T_{11}, T_{21}) and (T_{12}, T_{22}) be independent and identically distributed random vectors having joint distribution F in (4.3). For this setting, the definition of Kendall's tau given in Section 1.2.3 becomes the difference between the conditional probabilities of concordance and discordance given $T_{11} < \infty$ and $T_{12} < \infty$, that is,

$$\begin{aligned} \tau = {} & Pr((T_{11} - T_{12})(T_{21} - T_{22}) > 0 | T_{11} < \infty, T_{12} < \infty) \\ & - Pr((T_{11} - T_{12})(T_{21} - T_{22}) < 0 | T_{11} < \infty, T_{12} < \infty). \end{aligned} \tag{4.13}$$

By the definition and results in Section 1.2.3, τ is expressed in terms of the copula function C in (4.7) as follows:

$$\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1 \tag{4.14}$$

$$= 1 - 4 \int_0^1 \int_0^1 \frac{\partial C(u_1, u_2)}{\partial u_1} \frac{\partial C(u_1, u_2)}{\partial u_2} du_1 du_2. \tag{4.15}$$

Since $C(u_1, u_2) = p^{-1}C'(pu_1, u_2)$ for $0 \leq u_1 \leq 1$ and $0 \leq u_2 \leq 1$, from (4.14) and (4.15), τ can also be written in terms of the function C' in (4.3) as

$$\tau = \frac{4}{p^2}\int_0^1\int_0^p C'(u_1, u_2)dC'(u_1, u_2) - 1 \tag{4.16}$$

$$= 1 - \frac{4}{p^2}\int_0^1\int_0^p \frac{\partial C'(u_1, u_2)}{\partial u_1}\frac{\partial C'(u_1, u_2)}{\partial u_2}du_1du_2. \tag{4.17}$$

Consider the semi-survival model given in (4.5) and let $Pr(U_1 \leq u_1, U_2 > u_2) = \breve{C}'(u_1, 1-u_2)$ for $0 \leq u_1 \leq p$ and $0 \leq u_2 \leq 1$. Since $C'(u_1, u_2) = u_1 - \breve{C}'(u_1, 1-u_2)$, from (4.17) we obtain

$$\tau = \frac{4}{p^2}\int_0^1\int_0^p \frac{\partial \breve{C}'(u_1, u_2)}{\partial u_1}\frac{\partial \breve{C}'(u_1, u_2)}{\partial u_2}du_1du_2 - 1. \tag{4.18}$$

Similarly, when we consider the model given in (4.9), we let $Pr(U_1 \leq u_1, U_2 > u_2) = \breve{C}(u_1, 1-u_2)$ for $0 \leq u_1 \leq 1$ and $0 \leq u_2 \leq 1$. Since $C(u_1, u_2) = u_1 - \breve{C}(u_1, 1-u_2)$, from (4.15) we obtain

$$\tau = 4\int_0^1\int_0^1 \frac{\partial \breve{C}(u_1, u_2)}{\partial u_1}\frac{\partial \breve{C}(u_1, u_2)}{\partial u_2}du_1du_2 - 1. \tag{4.19}$$

When estimating Kendall's tau (4.13) parametrically, we can use existing formulas of Kendall's tau for copula families C such as in (1.30), (1.33) or (1.36). The parametric estimate of Kendall's tau is obtained by plugging in the parametric estimate of the dependence parameter(s).

When estimating Kendall's tau (4.13), as we only take into account the observed T_1 values, censoring may only affect the second survival time T_2. If we knew that $p = 1$ then to estimate Kendall's tau nonparametrically, a method presented in Wang and Wells (2000b) could be used. A nonparametric estimate of Kendall's tau, motivated by (4.14), is

$$\tilde{\tau} = 4\sum_{i=1}^{k}\sum_{j=1}^{k}\tilde{F}(t_{1(i)}, t_{2(j)})\tilde{F}(\Delta t_{1(i)}, \Delta t_{2(j)}) - 1 \tag{4.20}$$

where $\tilde{F}(t_1, t_2)$ is a nonparametric estimate of bivariate distribution function $F(t_1, t_2)$, $t_{1(1)} < t_{1(2)} < ... < t_{1(k)}$, $t_{2(1)} < t_{2(2)} < ... < t_{2(k)}$ are the ordered observations of t_1 and t_2 for the cases where t_1 is observed, $k = \sum_{i=1}^{n}\delta_{1i}$ and $\tilde{F}(\Delta t_{1(i)}, \Delta t_{2(j)}) = \tilde{F}(t_{1(i)}, t_{2(j)}) - \tilde{F}(t_{1(i)}, t_{2(j-1)}) - \tilde{F}(t_{1(i-1)}, t_{2(j)}) + \tilde{F}(t_{1(i-1)}, t_{2(j-1)})$ with $t_{1(0)} = 0$ and $t_{2(0)} = 0$.

If $F(t_1, t_2)$ is estimated nonparametrically by the method proposed by Lin et al. (1999), then $\tilde{F}(t_1, t_2) = \tilde{F}_1(t_1) - \tilde{H}(t_1, t_2)$ where $\tilde{F}_1(t_1)$ and $\tilde{H}(t_1, t_2)$ are given in (1.56) and (1.55),

respectively, for $t_1 + t_2 \leq C_{\max}$. After estimating the Kendall's tau by calculating (4.20), as described in Genest and Rivest (1993) and Wang and Wells (2000a), a nonparametric estimate of the dependence parameter α in the copula function C given in (4.7) can be obtained by equating $\tilde{\tau}$ to τ in (4.14) for the given copula function C if there is a one-to-one relationship between τ and α.

However, we never know for sure whether $p = 1$ or $p < 1$ and there will almost always be censoring in this setting. Unless all individuals have T_1 observed, this nonparametric estimation procedure cannot be used. Hence, another approach to model bivariate sequential data is presented in Section 4.4 in which we consider only individuals for whom $T_1 \leq Q$, with some suitably chosen Q.

4.2 Semiparametric Estimation Methods

4.2.1 A Two-Stage Procedure

We consider models in which the copula functions C in (4.11) or (4.12) are specified parametrically as $C_\alpha(u_1, u_2)$. In this section we give a procedure for estimating $F_0(t_1)$ and $F_2(t_2)$ nonparametrically, while obtaining parametric estimates of p and α. The distribution function $F_1(t_1)$ is first estimated by a Kaplan-Meier estimate as in Shih and Louis (1995). However, the nonparametric estimation of $F_2(t_2)$ is performed with the assumed copula family. The nonparametric estimate of $F_2(t_2)$ and the estimate of α are found simultaneously. The approach is as follows.

Consider the distribution function copula form (4.11). When p and $F_0(t_1)$ are fixed, the likelihood function (4.8) is proportional to the following, written in terms of $C_\alpha(F_0(t_1), F_2(t_2))$,

$$L = \prod_{i=1}^{n} \left[\frac{\partial^2 C_\alpha(F_0(t_{1i}), F_2(t_{2i}))}{\partial F_0(t_{1i}) \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \left[1 - \frac{\partial C_\alpha(F_0(t_{1i}), F_2(t_{2i}))}{\partial F_0(t_{1i})} \right]^{\delta_{1i}(1-\delta_{2i})}. \tag{4.21}$$

Note that α and $F_2(t_2)$ can only be estimated by using cases with T_1 observed (i.e., uncensored) since otherwise T_2 is not seen. Estimates of both $F_0(t_1)$ and p are needed for this purpose, but only $F_1(t_1) = pF_0(t_1)$ is estimable nonparametrically from observations on T_1, using Kaplan-Meier. In settings where it is felt that $p < 1$, we are forced to estimate p in an ad hoc way. Note that the basis for assuming $p < 1$ rests with background knowledge, and cannot be validated solely from the observed data. Evidence for $p < 1$ would be convincing only if $\hat{F}_1(t_1)$ has levelled off beyond some value $t_1^* < C_{\max}$. In that case, we will adopt the estimate $\hat{p} = \hat{F}_1(t_1^*)$ and then $\hat{F}_0(t_1) = \hat{p}^{-1}\hat{F}_1(t_1)$. In cases where $\hat{F}_1(t_1)$ is still increasing slightly up to $C_{\max}$, we could adopt a $\hat{p}$ that is a little larger than $\hat{F}_1(C_{\max})$.

In both cases we must depend on background knowledge to motivate the estimate $\hat{p} < 1$; there is nothing in the data that can guarantee that $p < 1$.

Thus, F_1 is estimated first by its Kaplan-Meier estimate, $\hat{F}_1$, and a plausible $\hat{p}$ is selected based on $\hat{F}_1$ and on background knowledge. When p and $F_0(t_{1i}) = F_1(t_{1i})/p$ are replaced by $\hat{p}$ and $\hat{F}_0(t_{1i}) = \hat{p}^{-1}\hat{F}_1(t_{1i})$ in (4.21), it becomes proportional to

$$L_s = \prod_{i=1}^{n} \left[\frac{\partial^2 C_\alpha(\hat{F}_0(t_{1i}), F_2(t_{2i}))}{\partial \hat{F}_0(t_{1i}) \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \left[1 - \frac{\partial C_\alpha(\hat{F}_0(t_{1i}), F_2(t_{2i}))}{\partial \hat{F}_0(t_{1i})} \right]^{\delta_{1i}(1-\delta_{2i})} . \tag{4.22}$$

Next, (4.22) is maximized with respect to F_2 and α. To do this we assume that $\hat{F}_2$ has jumps only at observed (i.e., uncensored) times t_2, so that the maximization problem becomes essentially parametric. It is convenient to use a discrete hazard parametrization as in Chapter 3, so we define $\lambda_1^* = F_2(t_{2(1)}^*)$ and $\lambda_l^* = \left(F_2(t_{2(l)}^*) - F_2(t_{2(l-1)}^*)\right) / \left(1 - F_2(t_{2(l-1)}^*)\right)$ for $l = 2, ..., k$ where $t_{2(1)}^* < t_{2(2)}^* < .... < t_{2(k)}^*$ are the distinct observed t_{2i}'s with $\delta_{2i} = 1$, for $i = 1, ..., n$. The likelihood (4.22) can be reexpressed by letting $\hat{u}_{1i} = \hat{F}_0(t_{1i})$ and defining $C_\alpha^{(1,2)}(u_1, u_2) = \partial^2 C_\alpha(u_1, u_2)/\partial u_1 \partial u_2$ and $C_\alpha^{(1)}(u_1, u_2) = \partial C_\alpha(u_1, u_2)/\partial u_1$, giving

$$L_s(\lambda^*, \alpha) = \prod_{i:\delta_{1i}=1} \left\{ C_\alpha^{(1,2)}(\hat{u}_{1i}, F_2(t_{2i})) \lambda_{l(t_{2i})} [1 - F_2(t_{2i}^-)] \right\}^{\delta_{2i}} \left\{ 1 - C_\alpha^{(1)}(\hat{u}_{1i}, F_2(t_{2i})) \right\}^{1-\delta_{2i}} , \tag{4.23}$$

where $F_2(t_{2i}) = 1 - \prod_{l:t_{2(l)}^* \le t_{2i}} (1 - \lambda_l^*)$, $F_2(t_{2i}^-) = 1 - \prod_{l:t_{2(l)}^* < t_{2i}} (1 - \lambda_l^*)$ and where for cases with $\delta_{2i} = 1$, $\lambda_{l(t_{2i})}$ is the corresponding λ_l^* where $l(t_{2i}) = l : t_{2(l)}^* = t_{2i}$. The logarithm of (4.23) is conveniently maximized using general purpose optimizers. We use the R function nlm in examples below.

Similarly, when we consider the semi-survival copula form $C_\alpha(F_0(t_1), S_2(t_2))$ given in (4.12), we define the parametrization $\lambda_1^* = 1 - S_2(t_{2(1)}^*)$ and $\lambda_l^* = 1 - S_2(t_{2(l)}^*)/S_2(t_{2(l-1)}^*)$ for $l = 2, ..., k$ and maximize the logarithm of the likelihood function

$$L_s(\lambda^*, \alpha) = \prod_{i:\delta_{1i}=1} \left\{ -C_\alpha^{(1,2)}(\hat{u}_{1i}, S_2(t_{2i})) \lambda_{l(t_{2i})} S_2(t_{2i}^-) \right\}^{\delta_{2i}} \left\{ C_\alpha^{(1)}(\hat{u}_{1i}, S_2(t_{2i})) \right\}^{1-\delta_{2i}} , \tag{4.24}$$

where $S_2(t_{2i}) = \prod_{l:t_{2(l)}^* \le t_{2i}} (1 - \lambda_l^*)$ and $S_2(t_{2i}^-) = \prod_{l:t_{2(l)}^* < t_{2i}} (1 - \lambda_l^*)$.

A nonparametric bootstrap procedure can be used to estimate the variance of the estimates of α, $F_2(t_2)$ and $S_2(t_2)$, and provide confidence intervals. To reflect all of the sampling variation present in the data, we reestimate all of the quantities $\hat{p}, \hat{\alpha}, \hat{F}_0$ and $\hat{F}_2$ for each bootstrap sample.

We remark that Shih and Louis (1995) were able to show asymptotic normality and obtain a variance estimate for $\hat{\alpha}$ for their two-stage procedure for parallel bivariate survival

times. This was reasonably straightforward since Kaplan-Meier estimates were inserted for both F_1 and F_2 at the second stage. The sequential case is much more complex, since both α and F_2 are estimated in stage 2, and no theoretical development of asymptotic results has yet been obtained. Properties of the estimators are studied by simulation in Section 4.3.

In Section 4.6 we also consider goodness of fit tests for copulas, based on embedding a copula family within a larger family as in Section 3.2. In particular, if $C_{\alpha_1,\alpha_2}(u_1, u_2)$ is a two-parameter family of copulas, we may test $H_0 : \alpha_2 = \alpha_{20}$ by considering a semiparametric pseudolikelihood ratio statistic, written in terms of p, F_0, F_2, α_1 and α_2 as

$$\Lambda_{s2}(\alpha_{20}) = 2\log L_s(\hat{p}, \hat{F}_0, \hat{F}_2, \hat{\alpha}_1, \hat{\alpha}_2) - 2\log L_s(\hat{p}, \hat{F}_0, \hat{F}_2(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}), \alpha_{20}),$$

where $\hat{F}_2(\alpha_{20})$ and $\hat{\alpha}_1(\alpha_{20})$ are obtained by maximizing (4.22) with α_2 fixed at α_{20}. In that case we obtain p-values by a semiparametric bootstrap procedure designed to respect the null hypothesis. The bootstrap procedure is similar to the one in Section 3.2.1 but it must follow the properties of sequential data.

4.2.2 Semiparametric Maximum Likelihood

In this case, nonparametric estimates of $F_0(t_1)$ and $F_2(t_2)$ and parametric estimates of α are obtained simultaneously. After estimating p as in the previous section, the likelihood function (4.8) for the copula model (4.7) is maximized with respect to F_0, F_2 and α by assuming that the estimates of F_0 and F_2 have jumps only at observed times t_1 and t_2, respectively. When we use a discrete hazard reparametrization, as in Section 4.2.1 we define

$$\lambda^*_{11} = F_0(t^*_{1(1)}),\ \lambda^*_{1l} = [F_0(t^*_{1(l)}) - F_0(t^*_{1(l-1)})]/[1 - F_0(t^*_{1(l-1)})]\ for\ l = 2, ..., k_1$$

and

$$\lambda^*_{21} = F_2(t^*_{2(1)}),\ \lambda^*_{2l} = [F_2(t^*_{2(l)}) - F_2(t^*_{2(l-1)})]/[1 - F_2(t^*_{2(l-1)})]\ for\ l = 2, ..., k_2$$

where $t^*_{1(1)} < t^*_{1(2)} < .... < t^*_{1(k_1)}$ and $t^*_{2(1)} < t^*_{2(2)} < .... < t^*_{2(k_2)}$ are distinct observed t_{1i}'s with $\delta_{1i} = 1$ and the distinct observed t_{2i}'s with $\delta_{2i} = 1$, respectively, for $i = 1, ..., n$ and $k_2 \le k_1$. The likelihood (4.8) can be reexpressed by defining $C^{(1,2)}_\alpha(u_1, u_2) = \partial^2 C_\alpha(u_1, u_2)/\partial u_1 \partial u_2$ and $C^{(1)}_\alpha(u_1, u_2) = \partial C_\alpha(u_1, u_2)/\partial u_1$, giving

$$L(\alpha, p, \lambda^*_1, \lambda^*_2) = \prod_{i=1}^{n} \left[p\lambda_{l_1(t_{1i})}(1 - F_0(t^-_{1i}))\right]^{\delta_{1i}} \left[\lambda_{l_2(t_{2i})}(1 - F_2(t^-_{2i}))C^{(1,2)}_\alpha(F_0(t_{1i}), F_2(t_{2i}))\right]^{\delta_{1i}\delta_{2i}}$$
$$\left[1 - C^{(1)}_\alpha(F_0(t_{1i}), F_2(t_{2i}))\right]^{\delta_{1i}(1-\delta_{2i})} \left[1 - pF_0(t_{1i})\right]^{1-\delta_{1i}}, \qquad (4.25)$$

where $F_0(t_{1i}) = 1 - \prod_{t:t^*_{1(l)} \le t_{1i}}(1 - \lambda^*_{1l})$, $F_0(t^-_{1i}) = 1 - \prod_{t:t^*_{1(l)} < t_{1i}}(1 - \lambda^*_{1l})$, $F_2(t_{2i}) = 1 - \prod_{t:t^*_{2(l)} \le t_{2i}}(1-\lambda^*_{2l})$, $F_2(t^-_{2i}) = 1-\prod_{t:t^*_{2(l)} < t_{2i}}(1-\lambda^*_{2l})$ and where for cases with $\delta_{1i} = 1$, $\lambda_{l_1(t_{1i})}$ is the corresponding λ^*_{1l} where $l_1(t_{1i}) = l : t_{1i} = t^*_{1(l)}$, and for cases with $\delta_{2i} = 1$, $\lambda_{l_2(t_{2i})}$ is the corresponding λ^*_{2l} where $l_2(t_{2i}) = l : t_{2i} = t^*_{2(l)}$. The estimates of the vectors λ^*_1, λ^*_2 and α are obtained by maximizing the logarithm of (4.25) with general purpose optimizers, where p is replaced by a plausible estimate $\hat{p}$. Software may run into problems if n is too large. However, when the R function nlm is used, for instance, it is observed that even when $n = 2000$ and there is 25% censoring for T_1 and 45% censoring for T_2, it is feasible to obtain the estimates. Once again we use the bootstrap for variance estimation.

As for the two-stage case in Section 4.2.1, there is currently no rigorous asymptotic theory for the estimators here. However, standard asymptotic normality and consistency seem plausible and are supported by simulation results. In the simpler case of parallel bivariate survival times, Chen et al. (2006) have recently shown that a sieve-based procedure gives estimators that are asymptotically normal and semiparametric efficient and under a specific model Li et al. (2008) have shown that the semiparametric maximum likelihood estimators of the association parameter and marginal survivals are consistent, asymptotically normal and semiparametric efficient.

Tests of fit for copula models can also be carried out as described at the end of the preceding section. In this case, the semiparametric likelihood ratio statistic is of the form

$$\Lambda_{s1}(\alpha_{20}) = 2\log L(\hat{p}, \hat{F}_0, \hat{F}_2, \hat{\alpha}_1, \hat{\alpha}_2) - 2\log L(\hat{p}, \hat{F}_0(\alpha_{20}), \hat{F}_2(\alpha_{20}), \hat{\alpha}_1(\alpha_{20}), \alpha_{20}),$$

where $\hat{F}_0(\alpha_{20})$, $\hat{F}_2(\alpha_{20})$ and $\hat{\alpha}_1(\alpha_{20})$ are obtained by maximizing (4.25) with α_2 fixed at α_{20}. We obtain p-values via a bootstrap procedure similar to the one in Section 3.2.1.

4.3 Simulation Study

A simulation study was conducted to study the performance of the two-stage semiparametric estimation and semiparametric maximum likelihood estimation procedures introduced in Section 4.2 and to compare them with a nonparametric estimation procedure (Lin et al., 1999) described in Section 1.4.2 and flexible parametric maximum likelihood estimation with piecewise constant specification for baseline hazard functions (He and Lawless, 2003). We generated 500 random bivariate survival time samples for each of sizes $n = 50$ and 100 from the Clayton copula model for (4.7), with $p = 1$, with two degrees of association represented by Kendall's tau values of $\tau = 0.4$ and 0.7. The marginal distributions of T_1 and T_2 were taken as Exponential with a unit scale parameter and Weibull with a unit scale parameter and shape parameter 2, respectively. The censoring times C_i were independently generated from the uniform distribution over (0,4) so that about 25% of T_1 and

45% of T_2 survival times were censored. Note that when T_1 is censored, T_2 is censored at 0, or unobserved.

Tables 4.1 and 4.2 show the empirical means and standard deviations of the semiparametric (one- and two-stage) estimators and of nonparametric and piecewise constant model maximum likelihood estimators of the conditional probability $Pr(T_2 > t_2|T_1 \leq t_1)$. We show results for $t_2 = 0.4724,\ 0.7147,\ 0.9572,\ 1.2686$ corresponding to marginal survival probabilities for T_2 of $0.8,\ 0.6,\ 0.4,\ 0.2$, and $t_1 = 0.5108,\ 1.6094$, corresponding to marginal survival probabilities for T_1 of $0.6,\ 0.2$. For the piecewise constant hazards approach, the time scales for T_1 and T_2 are divided into 8 pieces, with cut points 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 for both T_1 and T_2. Table 4.3 shows empirical means and standard deviations of the corresponding estimates for $F_2(t_2)$ and $\log \phi$. The nonparametric method does not estimate $F_2(t_2)$, hence results are given for only the semiparametric and piecewise constant methods.

Tables 4.1 and 4.3 indicate that when the assumed copula model is correct, semiparametric maximum likelihood and two-stage semiparametric estimators of $Pr(T_2 > t_2|T_1 \leq t_1)$ and $F_2(t_2)$ have little bias. For sample size $n = 50$, the semiparametric maximum likelihood estimator of $Pr(T_2 > t_2|T_1 \leq t_1)$ has generally smaller bias than the two-stage semiparametric one; but on the other hand, the two-stage semiparametric estimator of $F_2(t_2)$ has slightly smaller bias than the semiparametric maximum likelihood estimator. The bias of the flexible parametric maximum likelihood estimators of $Pr(T_2 > t_2|T_1 \leq t_1)$ or $F_2(t_2)$ depends on the number of pieces and cut points for the time scales for T_1 and T_2. The biases are fairly small in most cases, with occasional exceptions. It is known that well chosen flexible procedures mainly have bias problems near the ends of the range for the functions being estimated. The nonparametric estimator of $Pr(T_2 > t_2|T_1 \leq t_1)$ is asymptotically unbiased but note that for T_2 all that is nonparametrically estimable are these conditional probabilies for $t_1 + t_2 \leq C_{\max}$.

From Table 4.2, it is observed that semiparametric estimators of $Pr(T_2 > t_2|T_1 \leq t_1)$ are more efficient than the nonparametric estimators. The efficiency of the semiparametric estimators appear to be very similar when there is moderate association ($\tau = 0.4$), however, when there is strong association ($\tau = 0.7$), semiparametric maximum likelihood estimators seem to be slightly more efficient than two-stage estimators for small sample size. For estimation of $F_2(t_2)$ displayed in Table 4.3, the two semiparametric procedures are similar. For the association parameter ϕ, the two-stage estimator shows a little more bias than the two maximum likelihood estimators when $\tau = 0.4$. The piecewise constant model gives an efficient estimator of ϕ with little bias in the scenarios examined here.

			$\tau = 0.4$				$\tau = 0.7$			
			t_2				t_2			
n	t_1		0.4724	0.7147	0.9572	1.2686	0.4724	0.7147	0.9572	1.2686
50	0.5108	(a)	0.584	0.330	0.173	0.070	0.504	0.137	0.027	0.005
		(b)	0.578	0.320	0.168	0.068	0.511	0.136	0.028	0.005
		(c)	0.565	0.306	0.157	0.062	0.524	0.144	0.031	0.007
		(d)	0.584	0.324	0.152	0.065	0.514	0.143	0.026	0.007
		(e)	0.585	0.333	0.176	0.072	0.507	0.138	0.028	0.004
	1.6094	(a)	0.757	0.535	0.336	0.158	0.750	0.503	0.275	0.101
		(b)	0.753	0.530	0.334	0.155	0.754	0.506	0.281	0.102
		(c)	0.750	0.527	0.332	0.154	0.759	0.505	0.279	0.102
		(d)	0.756	0.531	0.308	0.151	0.754	0.504	0.257	0.106
		(e)	0.754	0.529	0.330	0.158	0.751	0.499	0.274	0.097
100	0.5108	(a)	0.584	0.330	0.173	0.070	0.504	0.137	0.027	0.005
		(b)	0.584	0.329	0.174	0.070	0.500	0.133	0.026	0.005
		(c)	0.575	0.320	0.167	0.067	0.502	0.136	0.027	0.006
		(d)	0.584	0.331	0.157	0.067	0.512	0.147	0.027	0.007
		(e)	0.584	0.332	0.179	0.071	0.503	0.137	0.027	0.005
	1.6094	(a)	0.757	0.535	0.336	0.158	0.750	0.503	0.275	0.101
		(b)	0.759	0.539	0.342	0.159	0.749	0.503	0.278	0.104
		(c)	0.754	0.534	0.338	0.157	0.750	0.503	0.277	0.103
		(d)	0.756	0.536	0.312	0.152	0.757	0.511	0.263	0.108
		(e)	0.758	0.536	0.338	0.155	0.753	0.503	0.278	0.104

Table 4.1: (a) True values and empirical means of (b) semiparametric maximum likelihood, (c) two-stage semiparametric, (d) flexible maximum likelihood and (e) nonparametric estimates of $Pr(T_2 > t_2 | T_1 \leq t_1)$.

			$\tau = 0.4$				$\tau = 0.7$			
			t_2				t_2			
n	t_1		0.4724	0.7147	0.9572	1.2686	0.4724	0.7147	0.9572	1.2686
50	0.5108	(a)	0.106	0.097	0.074	0.043	0.114	0.070	0.026	0.007
		(b)	0.106	0.095	0.071	0.040	0.136	0.091	0.041	0.020
		(c)	0.096	0.091	0.057	0.033	0.096	0.064	0.020	0.006
		(d)	0.128	0.128	0.107	0.075	0.131	0.091	0.046	0.020
	1.6094	(a)	0.073	0.089	0.092	0.074	0.073	0.088	0.088	0.068
		(b)	0.073	0.089	0.092	0.074	0.082	0.095	0.090	0.067
		(c)	0.068	0.086	0.071	0.055	0.063	0.078	0.065	0.046
		(d)	0.091	0.104	0.106	0.087	0.092	0.109	0.103	0.077
100	0.5108	(a)	0.077	0.066	0.047	0.030	0.080	0.046	0.014	0.005
		(b)	0.077	0.065	0.045	0.028	0.082	0.047	0.015	0.005
		(c)	0.070	0.060	0.037	0.022	0.071	0.044	0.013	0.004
		(d)	0.091	0.089	0.073	0.053	0.094	0.064	0.032	0.014
	1.6094	(a)	0.050	0.056	0.056	0.052	0.048	0.057	0.055	0.045
		(b)	0.050	0.056	0.056	0.051	0.048	0.058	0.057	0.045
		(c)	0.046	0.052	0.044	0.037	0.043	0.055	0.047	0.032
		(d)	0.064	0.070	0.069	0.061	0.063	0.068	0.068	0.057

Table 4.2: Empirical standard deviations of (a) semiparametric maximum likelihood, (b) two-stage semiparametric, (c) flexible maximum likelihood and (d) nonparametric estimates of $Pr(T_2 > t_2|T_1 \leq t_1)$ over 500 simulations.

		$\tau = 0.4$					$\tau = 0.7$				
					t_2					t_2	
n		$\log\phi$	0.4724	0.7147	0.9572	1.2686	$\log\phi$	0.4724	0.7147	0.9572	1.2686
	(a)	0.288	0.2	0.4	0.6	0.8	1.540	0.2	0.4	0.6	0.8
50	(b)	0.344	0.213	0.418	0.614	0.810	1.655	0.200	0.405	0.604	0.811
		(0.443)	(0.064)	(0.085)	(0.097)	(0.086)	(0.340)	(0.061)	(0.075)	(0.091)	(0.102)
	(c)	0.426	0.206	0.404	0.598	0.798	1.606	0.194	0.399	0.599	0.802
		(0.400)	(0.061)	(0.083)	(0.099)	(0.090)	(0.786)	(0.067)	(0.081)	(0.094)	(0.103)
	(d)	0.329	0.202	0.405	0.628	0.806	1.582	0.198	0.402	0.621	0.794
		(0.378)	(0.059)	(0.081)	(0.078)	(0.068)	(0.296)	(0.051)	(0.070)	(0.078)	(0.078)
100	(b)	0.302	0.203	0.403	0.600	0.803	1.606	0.208	0.413	0.615	0.813
		(0.281)	(0.043)	(0.054)	(0.060)	(0.061)	(0.205)	(0.040)	(0.053)	(0.060)	(0.068)
	(c)	0.361	0.203	0.399	0.594	0.798	1.581	0.201	0.401	0.600	0.799
		(0.263)	(0.042)	(0.054)	(0.061)	(0.063)	(0.201)	(0.040)	(0.053)	(0.063)	(0.072)
	(d)	0.307	0.202	0.400	0.625	0.807	1.498	0.196	0.396	0.617	0.798
		(0.247)	(0.040)	(0.051)	(0.049)	(0.046)	(0.185)	(0.036)	(0.052)	(0.057)	(0.055)

Table 4.3: (a) True values and empirical means of (b) semiparametric maximum likelihood, (c) two-stage semiparametric, (d) flexible maximum likelihood estimates of ϕ and $F_2(t_2)$ when $t_2 = 0.4724,\ 0.7147,\ 0.9572$ and 1.2686. The corresponding empirical standard deviations are given in paranthesis.

4.4 Another Approach to Model Bivariate Sequential Data

It is of interest to estimate a copula parameter without assumptions on the marginal distributions. Since the marginal distribution of T_2 is nonparametrically inestimable and we can only estimate certain values $Pr(T_1 \leq t_1, T_2 \leq t_2)$ or $Pr(T_2 \leq t_2 | T_1 \leq t_1)$, it is essential to restrict our attention to $T_1 \in [0, Q]$ where $Q < C_{\max}$. If Q is a given value then we can estimate $Pr(T_2 \leq t_2 | T_1 \leq Q)$ nonparametrically for $0 \leq t_2 \leq C_{\max} - Q$. Another approach to model bivariate sequential data is to use a copula model for the truncated distribution $Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 \leq Q)$ where Q is some selected value. Thus, consider models of the form

$$F_Q(t_1, t_2) = Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 \leq Q) = C^Q(F_{1Q}(t_1), F_{2Q}(t_2)) \tag{4.26}$$

where $F_{1Q}(t_1) = Pr(T_1 \leq t_1 | T_1 \leq Q)$ and $F_{2Q}(t_2) = Pr(T_2 \leq t_2 | T_1 \leq Q)$, where C^Q is a copula. Other forms such as the semi-survival form

$$H_Q(t_1, t_2) = Pr(T_1 \leq t_1, T_2 > t_2 | T_1 \leq Q) = C^Q(F_{1Q}(t_1), S_{2Q}(t_2)) \tag{4.27}$$

where $S_{2Q}(t_2) = 1 - F_{2Q}(t_2)$ can also be considered.

Note that $C^Q(F_{1Q}(t_1), F_{2Q}(t_2))$ can be written in terms of $C'(F_1(t_1), F_2(t_2))$ and $C(F_0(t_1), F_2(t_2))$ defined in (4.3) and (4.7), respectively, as follows: For $0 \leq t_1 < Q$ and $0 \leq t_2 < \infty$,

$$C^Q(F_{1Q}(t_1), F_{2Q}(t_2)) = \frac{1}{p_Q} C'(F_1(t_1), F_2(t_2)) \tag{4.28}$$

and by (4.11),

$$C^Q(F_{1Q}(t_1), F_{2Q}(t_2)) = \frac{p}{p_Q} C(F_0(t_1), F_2(t_2)) \tag{4.29}$$

where $p_Q = Pr(T_1 < Q) = F_1(Q) = pF_0(Q)$. Note also that

$$F_{1Q}(t_1) = \frac{1}{p_Q} F_1(t_1) = \frac{F_0(t_1)}{F_0(Q)} \tag{4.30}$$

and

$$F_{2Q}(t_2) = \frac{1}{p_Q} C'(p_Q, F_2(t_2)) = \frac{1}{F_0(Q)} C(F_0(Q), F_2(t_2)) \tag{4.31}$$

Hence, it is also possible to fit the models (4.26) and (4.27). One limitation of these models is that it is harder to interpret $F_{2Q}(t_2)$ or $S_{2Q}(t_2)$ than $Pr(T_2 \leq t_2 | T_1 = t_1)$ or $Pr(T_2 > t_2 | T_1 = t_1)$. In addition, if copula $C(F_0(t_1), F_2(t_2))$ has a specific parametric form then this restricts the form of $F_{2Q}(t_2)$ due to (4.31) and, therefore, $C^Q(F_{1Q}(t_1), F_{2Q}(t_2))$ is

not a regular form. Thus when using (4.26) or (4.27), we prefer to choose a familiar copula form for C^Q; then C' defined by (4.28) or C defined by (4.29) is not a copula. However, it turns out that if C is of the Clayton form, so is C^Q.

Note that C^Q has the same copula form as C iff C is the independent or the Clayton copula. To show this, first assume C is an Archimedean copula (1.27). Then, (4.29) becomes

$$C^Q(F_{1Q}(t_1), F_{2Q}(t_2)) = \frac{p}{p_Q}\varphi^{-1}[\varphi(F_0(t_1)) + \varphi(F_2(t_2))] \tag{4.32}$$

and from (4.31) we obtain

$$F_2(t_2) = \varphi^{-1}\left[\varphi\left(\frac{p_Q}{p}F_{2Q}(t_2)\right) - \varphi\left(\frac{p_Q}{p}\right)\right] \tag{4.33}$$

When $F_0(t_1)$ and $F_2(t_2)$ are replaced by $\frac{p_Q}{p}F_{1Q}(t_1)$ and (4.33), respectively, (4.32) is written as

$$C^Q(F_{1Q}(t_1), F_{2Q}(t_2)) = \frac{p}{p_Q}\varphi^{-1}\left[\varphi\left(\frac{p_Q}{p}F_{1Q}(t_1)\right) + \varphi\left(\frac{p_Q}{p}F_{2Q}(t_2)\right) - \varphi\left(\frac{p_Q}{p}\right)\right] \tag{4.34}$$

Sungur (2002) showed that the general solution of the functional equation given in (4.34) is $\varphi(v) = \gamma \log v$ or $\varphi(v) = \rho(v^\gamma - 1)$ with some constants $\gamma,\ \rho$ and he proved that if C is an Archimedean copula, it is truncation dependence invariant iff C is the independent copula or the Clayton copula. Oakes (2005) strengthened this result and proved that the independent copula and the Clayton copula are the only ones which are truncation dependence invariant in all classes of copulas.

4.5 Two-Stage Semiparametric Estimation for Truncated Models

A possible approach to estimate the vector of dependence parameters α in the copula function C_α^Q in (4.26) or (4.27) is to apply a similar method to that of Shih and Louis (1995) which is described in Section 1.4.1. In the first stage, F_{1Q} and F_{2Q} are estimated nonparametrically, and in the second stage, α is estimated. The details of the estimation procedure is as follows.

Consider the models given in (4.26) and (4.27) and suppose F_{1Q} and F_{2Q} are known. Then the likelihood function is written in terms of $C_\alpha^Q(F_{1Q}(t_1), F_{2Q}(t_2))$ as follows:

$$L = \prod_{\substack{i=1 \\ t_{1i}\le Q}}^{n} p_Q^{\delta_{1i}}\left[\frac{\partial^2 C_\alpha^Q(F_{1Q}(t_{1i}), F_{2Q}(t_{2i}))}{\partial F_{1Q}(t_{1i})\partial F_{2Q}(t_{2i})}\right]^{\delta_{1i}\delta_{2i}}\left[1 - \frac{\partial C_\alpha^Q(F_{1Q}(t_{1i}), F_{2Q}(t_{2i}))}{\partial F_{1Q}(t_{1i})}\right]^{\delta_{1i}(1-\delta_{2i})}$$
$$\times (1 - p_Q F_{1Q}(t_{1i}))^{1-\delta_{1i}} \tag{4.35}$$

and it is written in terms of the semi-survival model $\breve{C}_\alpha^Q(F_{1Q}(t_1), S_{2Q}(t_2))$ as follows:

$$L = \prod_{\substack{i=1 \\ t_{1i} \le Q}}^{n} p_Q^{\delta_{1i}} \left[\frac{\partial^2 \breve{C}_\alpha^Q(F_{1Q}(t_{1i}), S_{2Q}(t_{2i}))}{\partial F_{1Q}(t_{1i}) \partial S_{2Q}(t_{2i})} \right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial \breve{C}_\alpha^Q(F_{1Q}(t_{1i}), S_{2Q}(t_{2i}))}{\partial F_{1Q}(t_{1i})} \right]^{\delta_{1i}(1-\delta_{2i})} \times (1 - p_Q F_{1Q}(t_{1i}))^{1-\delta_{1i}}. \tag{4.36}$$

Analogous to Shih and Louis (1995), in the first stage F_{1Q}, p_Q and F_{2Q} can be estimated nonparametrically by the methods proposed by Lin et al. (1999) or Schaubel and Cai (2004a) which are summarized in Section 1.4.2. Let $\tilde{F}_{1Q}(t_{1i})$, $\tilde{p}_Q$ and $\tilde{F}_{2Q}(t_{2i})$ be the nonparametric estimates of $F_{1Q}(t_{1i})$, p_Q and $F_{2Q}(t_{2i})$. Then, in the second stage, after replacing $F_{1Q}(t_{1i})$, p_Q and $F_{2Q}(t_{2i})$ in (4.35) with $\tilde{F}_{1Q}(t_{1i})$, $\tilde{p}_Q$ and $\tilde{F}_{2Q}(t_{2i})$, the semiparametric estimate of the vector of dependence parameters α is obtained by maximizing the likelihood function

$$L_s = \prod_{\substack{i=1 \\ t_{1i} \le Q}}^{n} \left[\frac{\partial^2 C_\alpha^Q(\tilde{F}_{1Q}(t_{1i}), \tilde{F}_{2Q}(t_{2i}))}{\partial \tilde{F}_{1Q}(t_{1i}) \partial \tilde{F}_{2Q}(t_{2i})} \right]^{\delta_{1i}\delta_{2i}} \left[1 - \frac{\partial C_\alpha^Q(\tilde{F}_{1Q}(t_{1i}), \tilde{F}_{2Q}(t_{2i}))}{\partial \tilde{F}_{1Q}(t_{1i})} \right]^{\delta_{1i}(1-\delta_{2i})}. \tag{4.37}$$

Similarly, if $\tilde{S}_{2Q}(t_{2i})$ is the nonparametric estimate of $S_{2Q}(t_{2i})$ then the semiparametric estimate of α is obtained from (4.36) by maximizing

$$L_s = \prod_{\substack{i=1 \\ t_{1i} \le Q}}^{n} \left[\frac{\partial^2 \breve{C}_\alpha^Q(\tilde{F}_{1Q}(t_{1i}), \tilde{S}_{2Q}(t_{2i}))}{\partial \tilde{F}_{1Q}(t_{1i}) \partial \tilde{S}_{2Q}(t_{2i})} \right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial \breve{C}_\alpha^Q(\tilde{F}_{1Q}(t_{1i}), \tilde{S}_{2Q}(t_{2i}))}{\partial \tilde{F}_{1Q}(t_{1i})} \right]^{\delta_{1i}(1-\delta_{2i})}. \tag{4.38}$$

In future work, properties of the two-stage semiparametric estimation procedure for the truncated distribution (4.26) or (4.27) will be investigated. The nonparametric bootstrap can presumably be used to obtain variance estimates or confidence intervals, but it may also be feasible to combine asymptotic results in Lin et al. (1999) and Schaubel and Cai (2004a) with ones in Shih and Louis (1995), in order to obtain an explicit variance estimate for the semiparametric estimate of α.

4.6 Colon Cancer Data

The data described in Section 1.1.2 consist of the time t_1 from study registration to cancer recurrence or censoring and if the cancer recurrence occurred, the time t_2 from cancer recurrence to death or censoring, for both the placebo and the levamisole plus fluorouracil therapy group. There were 315 patients in the placebo group and 304 patients in the

therapy group. By the end of the study, 177 patients in the placebo group had cancer recurrence, among whom 155 died, while in the therapy group 119 patients had cancer recurrence, among whom 108 died.

The Kaplan-Meier estimates for the survivor functions of T_1 in the two treatment groups suggest that the hazard function for recurrence in each becomes small for large t. Since it appears some subjects may be cured and never have a recurrence, a model where $F_1(\infty) = p < 1$ is useful. Here, a fraction $1 - p$ of patients is assumed to have no chance of disease recurrence. Lawless (2003, page 181) used a cure-rate model with distribution function (4.1). He showed that the log-logistic form for $F_0(t_1)$ (i.e., $F_0(t_1) = 1 - [1 + (t_1/\alpha_1)^{\beta_1}]^{-1}$) fits well for both treatment groups.

We will describe the use of the semiparametric estimation procedures of Section 4.2 for both estimation and model checking. We first fitted fully parametric models given by (4.11) to the data. We considered a log-logistic for $F_0(t_1)$ and log-logistic and Weibull forms for $F_2(t_2)$ (i.e., $F_2(t_2) = 1 - [1 + (t_2/\alpha_2)^{\beta_2}]^{-1}$ and $F_2(t_2) = 1 - \exp[-(t_2/\alpha_2)^{\beta_2}]$, respectively). We performed model checking by embedding some proposed copula families for C in an expanded family of copulas for these two cases. We considered the two-parameter copula family given in (1.40) where $u_1 = F_0(t_1)$ and $u_2 = F_2(t_2)$ and tested whether the proposed model Clayton (1.28) or Gumbel-Hougaard (1.31) represents the data adequately or not. For the control group, the maximum likelihood estimates of the parameters of the proposed and the expanded copula families, their standard errors and the maximized log-likelihood values of the corresponding model are given in Table 4.4 and 4.5 when the form of $F_2(t_2)$ is considered as log-logistic and Weibull, respectively. For the treatment group, the corresponding results can be seen in Table 4.6 and 4.7.

For the control and treatment groups, it is observed that the maximized log-likelihood values when the log-logistic distribution is assumed for $F_2(t_2)$ are higher than when the Weibull distribution is assumed. Hence, log-logistic distribution provides a better fit than the Weibull distribution. When testing whether C belongs to the Clayton family, under $H_0 : \theta = 1$, the log-likelihood ratio statistic $\Lambda_1(1) = 2(\ell(\hat{\beta}, \hat{\phi}, \hat{\theta}) - \ell(\hat{\beta}(\theta = 1), \hat{\phi}(\theta = 1), 1))$ has a limiting distribution with $Pr(\Lambda_1(1) \leq q) = 0.5 + 0.5Pr(\chi^2_{(1)} \leq q)$ where l is the logarithm of the likelihood function L in (4.8) and β is the vector of parameters including p and the parameters in the distribution functions $F_0(t_1)$ and $F_2(t_2)$. For the control group, it is concluded that there is a strong evidence against the Clayton model with p-value $0.5Pr(\chi^2_{(1)} \geq 15.215) \leq 5 \times 10^{-5}$ (when $F_2(t_2)$ is in the log-logistic form). When testing whether C belongs to the Gumbel-Hougaard family, under $H_0 : \phi = 0$, the log-likelihood ratio statistic $\Lambda_1(0) = 2(\ell(\hat{\beta}, \hat{\phi}, \hat{\theta}) - \ell(\hat{\beta}(\phi = 0), 0, \hat{\theta}(\phi = 0)))$ has a limiting distribution with $Pr(\Lambda_1(0) \leq q) = 0.5 + 0.5Pr(\chi^2_{(1)} \leq q)$. For the control group, it is concluded that there is no evidence against the Gumbel-Hougaard model. For the treatment group, there is no evidence against either model with p-values $0.5Pr(\chi^2_{(1)} \geq 2.466) = 0.058$ and $0.5Pr(\chi^2_{(1)} \geq 1.733) = 0.094$ when $H_0 : \theta = 1$ and $H_0 : \phi = 0$, respectively, are true, but

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{p}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{p}))$	$(se(\hat{\alpha}_1))$	$(se(\hat{\beta}_1))$	$(se(\hat{\alpha}_2))$	$(se(\hat{\beta}_2))$
Two-parameter	-2639.254	2×10^{-6}	1.421	0.610	426.240	1.566	462.403	1.434
		(2×10^{-6})	(0.112)	(0.033)	(40.647)	(0.124)	(43.691)	(0.102)
Clayton	-2646.861	0.306	1	0.609	429.741	1.592	436.552	1.494
		(0.116)	(-)	(0.033)	(40.820)	(0.125)	(38.099)	(0.104)
Gumbel-Hougaard	-2639.254	0	1.421	0.610	426.25	1.566	462.414	1.434
		(-)	(0.111)	(0.033)	(40.640)	(0.124)	(43.677)	(0.102)

Table 4.4: Maximum likelihood estimation results for the control group of colon cancer data when the model (4.7) is used and $F_2(t_2)$ is in log-logistic form.

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{p}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{p}))$	$(se(\hat{\alpha}_1))$	$(se(\hat{\beta}_1))$	$(se(\hat{\alpha}_2))$	$(se(\hat{\beta}_2))$
Two-parameter	-2642.636	2×10^{-5}	1.383	0.610	425.321	1.567	726.158	0.997
		(2×10^{-5})	(0.116)	(0.033)	(40.948)	(0.126)	(65.361)	(0.066)
Clayton	-2645.924	0.462	1	0.612	425.803	1.548	676.469	1.038
		(0.145)	(-)	(0.033)	(41.503)	(0.125)	(54.303)	(0.067)
Gumbel-Hougaard	-2642.636	0	1.383	0.610	425.318	1.567	726.154	0.997
		(-)	(0.114)	(0.033)	(40.888)	(0.126)	(65.225)	(0.066)

Table 4.5: Maximum likelihood estimation results for the control group of colon cancer data when the model (4.7) is used and $F_2(t_2)$ is in Weibull form.

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{p}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{p}))$	$(se(\hat{\alpha}_1))$	$(se(\hat{\beta}_1))$	$(se(\hat{\alpha}_2))$	$(se(\hat{\beta}_2))$
Two-parameter	-1833.210	0.210	1.290	0.429	491.675	1.658	321.173	1.382
		(0.177)	(0.202)	(0.032)	(53.013)	(0.167)	(37.637)	(0.122)
Clayton	-1834.444	0.454	1	0.426	489.437	1.712	309.660	1.452
		(0.138)	(-)	(0.032)	(50.856)	(0.163)	(33.556)	(0.116)
Gumbel-Hougaard	-1834.077	0	1.503	0.431	495.100	1.613	326.969	1.341
		(-)	(0.151)	(0.033)	(55.248)	(0.162)	(40.047)	(0.116)

Table 4.6: Maximum likelihood estimation results for the treatment group of colon cancer data when the model (4.7) is used and $F_2(t_2)$ is in log-logistic form.

Copula Model	Log-likelihood	$\hat{\phi}$	$\hat{\theta}$	$\hat{p}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$
		$(se(\hat{\phi}))$	$(se(\hat{\theta}))$	$(se(\hat{p}))$	$(se(\hat{\alpha}_1))$	$(se(\hat{\beta}_1))$	$(se(\hat{\alpha}_2))$	$(se(\hat{\beta}_2))$
Two-parameter	-1837.533	0.601	1.011	0.431	488.953	1.604	506.988	0.964
		(0.321)	(0.204)	(0.033)	(55.342)	(0.165)	(60.245)	(0.081)
Clayton	-1837.534	0.616	1	0.431	488.728	1.605	505.357	0.966
		(0.184)	(-)	(0.033)	(55.073)	(0.164)	(52.364)	(0.070)
Gumbel-Hougaard	-1840.405	0	1.440	0.433	501.371	1.596	550.794	0.891
		(-)	(0.163)	(0.033)	(58.229)	(0.168)	(68.226)	(0.070)

Table 4.7: Maximum likelihood estimation results for the treatment group of colon cancer data when the model (4.7) is used and $F_2(t_2)$ is in Weibull form.

the Gumbel-Hougaard has a slightly larger maximum likelihood and p-value.

The estimated marginal distributions for T_1 and T_2 as well as goodness of fit test just performed for the copula could be sensitive to misspecification of the parametric marginal distributions, so we now turn to semiparametric estimation. In Tables 4.8 and 4.9, the two-stage semiparametric and semiparametric maximum likelihood estimates of the parameters in the copula models and the maximized log-likelihood values of the corresponding model are shown. In both control and treatment groups, $\hat{p}$ is based on the Kaplan-Meier estimate of F_1, $\hat{F}_1$, and it is estimated as slightly larger than $\hat{F}_1(C_{\max})$. We chose $\hat{p} = 0.599$ for the control group and $\hat{p} = 0.405$ for the treatment group.

Values of the likelihood and pseudolikelihood ratio test statistics for testing the Clayton and Gumbel-Hougaard copula models obtained from Tables 4.4 - 4.9 are summarized in Table 4.10. When semiparametric likelihood and pseudolikelihood ratio statistics given in Secion 4.2 are used, we again reach the conclusion that there is no evidence against the Gumbel-Hougaard model and there is strong evidence against the Clayton model for the control group. For the semiparametric likelihood ratio statistic we estimated p-value by a semiparametric bootstrap procedure with 1000 samples, designed to respect the null hypothesis. The p-value estimated in this way for the Clayton model was 0. On the other hand, for the treatment group, although we again conclude that there is no evidence against either model, the Clayton model fits better according to semiparametric likelihood and pseudolikelihood ratio tests. The Gumbel-Hougaard model is not rejected with estimated p-value 0.065 for the semiparametric likelihood ratio statistic.

We next undertake further model checks. To assess the adequacy of the parametric log-logistic form of $F_2(t_2)$, we compare the parametric and semiparametric estimates in Figure 4.1. For the control group, the estimates are based on the Gumbel-Hougaard copula and for the treatment group, they are based on the Clayton copula. There is essentially no difference between the two semiparametric estimates but the parametric estimates depart substantially from the semiparametric ones for both groups. Therefore, the log-logistic distribution assumption for T_2 is questionable. Note that the nonparametric estimates $\hat{F}_2(t_2)$ for the two treatment groups in Figure 4.1 indicate the survival probability of patients who were on therapy is slightly lower at various times t_2 than for patients in the control group, suggesting that although treatment decreases the risk of recurrence, patients who experience a recurrence tend to survive slightly less long if they received the treatment. The same conclusion was reached by Lin et al. (1999) and He and Lawless (2003).

Since there can also be a misspecification of the copula family, we compare the semiparametric estimates of $Pr(T_2 > t_2|T_1 \leq t_1)$ with nonparametric estimates given in Lin et al. (1999) and Schaubel and Cai (2004a). Again, the semiparametric estimates are based on the Gumbel-Hougaard copula for the control group and the Clayton copula for the treatment group. In Figure 4.2, plots of the semiparametric and nonparametric estimates

	Two-Stage Semiparametric			Semiparametric Maximum Likelihood		
Copula Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$
Two-parameter	-805.230	9×10^{-6}	1.400	-1914.849	0	1.395
Clayton	-810.352	0.422	1	-1920.713	0.398	1
Gumbel-Hougaard	-805.230	0	1.400	-1914.849	0	1.395

Table 4.8: Two-stage semiparametric and semiparametric maximum likelihood estimation results for the control group of colon cancer data.

	Two-Stage Semiparametric			Semiparametric Maximum Likelihood		
Copula Model	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$	Log-likelihood	$\tilde{\phi}$	$\tilde{\theta}$
Two-parameter	-510.165	0.744	1	-1276.061	0.687	1
Clayton	-510.165	0.744	1	-1276.061	0.687	1
Gumbel-Hougaard	-514.651	0	1.377	-1278.011	0	1.519

Table 4.9: Two-stage semiparametric and semiparametric maximum likelihood estimation results for the treatment group of colon cancer data.

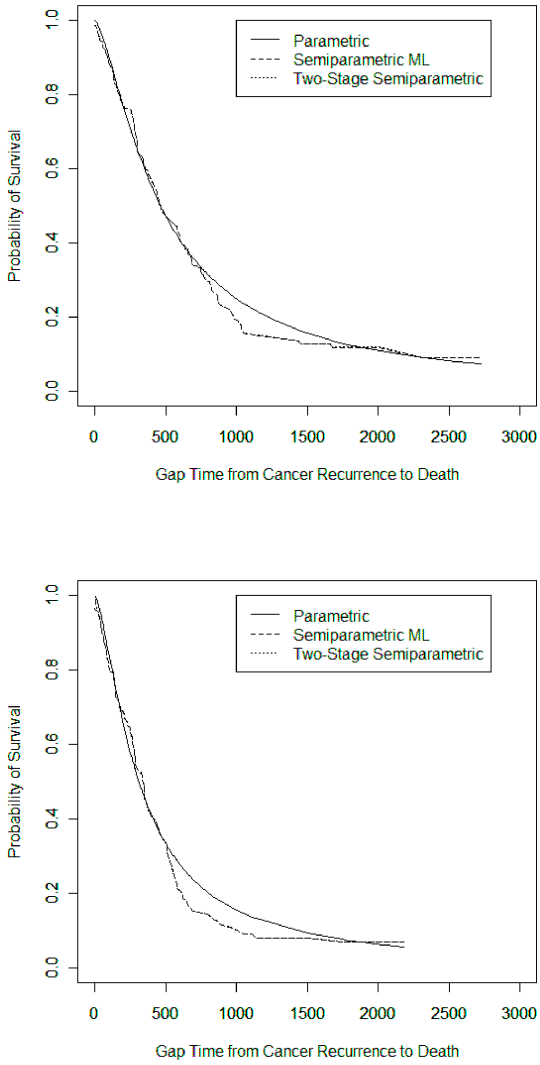


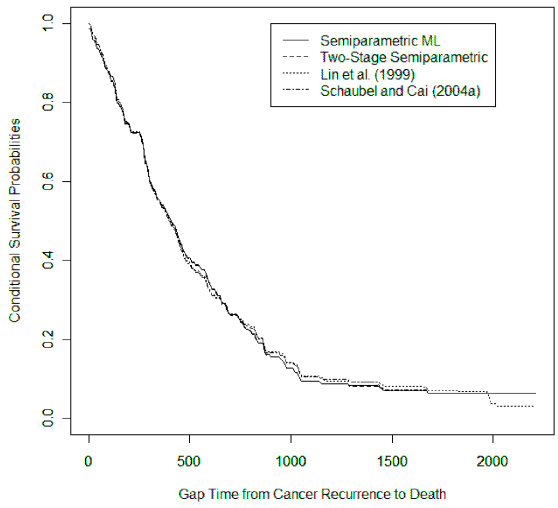
Figure 4.1: Parametric, semiparametric maximum likelihood and two-stage semiparametric estimates of $S_2(t_2) = 1 - F_2(t_2)$ for the control (top plot) and the treatment (bottom plot) groups.

	Control Group			
Null Copula Model	Log-Logistic F_2	Weibull F_2	2-SP	SPML
Clayton	15.215	6.576	10.244	11.728
Gumbel-Hougaard	0	0	0	0
	Treatment Group			
Null Copula Model	Log-logistic F_2	Weibull F_2	2-SP	SPML
Clayton	2.466	0.002	0	0
Gumbel-Hougaard	1.733	5.744	8.972	3.900

Table 4.10: Values of (pseudo)likelihood ratio test statistics for testing the Clayton and the Gumbel-Hougaard copula models when $F_2(t_2)$ has the log-logistic and Weibull forms and when it is nonparametrically estimated through the two-stage semiparametric (2-SP) and semiparametric maximum likelihood (SPML) estimation methods.

of conditional probability $Pr(T_2 > t_2|T_1 \leq 1000)$ for the control and treatment groups are shown. There is essentially no difference between the two semiparametric estimates and they are very close to the nonparametric estimates. Hence, we once again have no evidence against the Gumbel-Hougaard and the Clayton copula functions for the control and treatment groups, respectively.

Finally, we note that the copula models indicate a mild degree association between the survival times T_1 and T_2 for an individual. Kendall's tau in (4.14) can be estimated, for example, by plugging in the two-stage semiparametric estimate of θ in the Gumbel-Hougaard copula into (1.33) for the control group and it is obtained that $\tilde{\tau}_{control} = 0.286$. A standard error based on 100 bootstrap samples was found as 0.063. By comparison, the fully parametric model in Table 4.6 gives $\hat{\tau}_{control} = 0.296$ and a standard error of 0.055. For the treatment group Kendall's tau can be estimated by plugging in the two-stage semiparametric estimate of ϕ in the Clayton copula into (1.30) and it is obtained that $\tilde{\tau}_{trt} = 0.271$ with a standard error 0.060 based on 100 bootstrap samples.



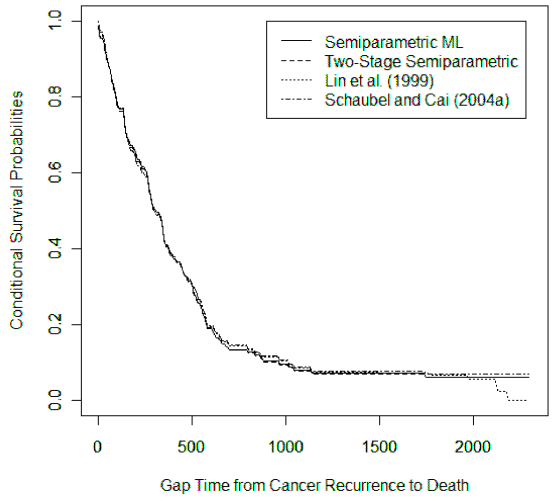


Figure 4.2: Semiparametric maximum likelihood, two-stage semiparametric and nonparametric (Lin et al., 1999; Schaubel and Cai, 2004a) estimates of $Pr(T_2 > t_2 | T_1 \leq 1000)$ for the control (top plot) and the treatment (bottom plot) groups.

Chapter 5

Summary and Further Research

5.1 Likelihood-Based Tests of Fit for Parametric Models

We examined tests for the adequacy of a copula-based bivariate survival time model, based on embedding a model in a larger copula family. In the fully parametric setting, where the marginal distributions have parametric specifications, both likelihood ratio test statistics and pseudolikelihood ratio test statistics arising from two-stage estimation (Shih and Louis, 1995) have asymptotic null distributions of chi-squared type, and p-values can be obtained via large sample approximation or by simulation. The tests have good power even when the expanded family does not include the alternative, and are asymptotically most powerful when it does include the alternative.

A natural question is how to select the expanded copula family. For the most commonly used single parameter models, such as the Clayton-Oakes, Gumbel-Hougaard and Frank copulas, there are two- or three-parameter copula families that include them (see Joe 1997, Sections 5.2 and 5.3; Genest et al., 1998) and we recommend their use. Similarly, Gaussian copulas can be embedded within t or skew-t families. However, an investigation of ways to obtain expanded models for an arbitrary copula family, as done with other smooth tests of fit (e.g. Pena, 1998) would be of interest.

5.2 Semiparametric Estimation for Parallel Clustered Data

We introduced semiparametric maximum likelihood estimation in which the copula parameter is estimated without assumptions on the marginal distributions by assuming the estimates of the marginal distribution functions have jumps only at observed times. It was shown that the semiparametric maximum likelihood estimation approach can also be used to fit copula models with proportional hazards margins for data with covariates. Furthermore, the two-stage estimation approach introduced by Shih and Louis (1995) was also extended to fit models with proportional hazards margins.

A simulation study was done to assess the performance of the semiparametric maximum likelihood estimator and two-stage semiparametric estimator (Shih and Louis, 1995) of the copula parameter. Two-stage semiparametric estimator is found to be as good as the semiparametric maximum likelihood estimator.

5.3 Likelihood-Based Tests of Fit for Semiparametric Models

The likelihood ratio approach can also be used with semiparametric models, as illustrated in Chapter 3. Methods for obtaining p-values for both semiparametric likelihood ratio and pseudolikelihood ratio tests were given. In simulation studies, it was observed that the semiparametric pseudolikelihood ratio test is almost as powerful as the parametric likelihood ratio and pseudolikelihood ratio tests while achieving robustness to the form of the marginal distributions.

Work on semiparametric maximum likelihood by Murphy and van der Vaart (2000) suggests that likelihood ratio tests and pseudolikelihood ratio tests about copula parameters might have chi-squared asymptotics similar to those for fully parametric models, for cases where the marginal distributions are non- or semi- parametrically specified. However, semiparametric maximum likelihood has been studied only in special situations, and in particular, has not been considered when covariates are present (see Li et al., 2008; Chen et al., 2006 and references therein). Moreover, cases where parameter values lie on the boundary of the parameter space have not been considered; these arise with many of our tests. The two-stage approach has also not been throughly investigated in the semiparametric setting, although Shih and Louis (1995) show that, under some conditions, regular asymptotics hold for the estimation of a copula parameter when the marginal distributions are estimated nonparametrically from censored data by Kaplan-Meier. The development of asymptotic theory for these settings poses challenging problems.

Simulation studies show that an adjustment for the semiparametric likelihood and pseudolikelihood ratio statistics is generally necessary to obtain that the distribution of the statistics is approximated by a chi-squared distribution. However, since it is hard to find the correction term for each copula model separately, especially under censoring, and since correction terms may depend on the unknown parameters and a copula parameter can be on the boundary, it is suggested to estimate p-values by using the proposed bootstrap procedures.

Finally, we have considered in our simulations the case where the bivariate lifetimes are observed in parallel. In some cases, lifetimes for an individual are observed in sequence (Lin et al., 1999; Visser, 1996). The tests for fully parametric models readily handle these other settings but semiparametric models require additional study. We remark that a flexible alternative to semiparametric estimation is to use weakly parametric models for the marginal distributions. He and Lawless (2003), for example, consider piecewise constant and spline models for marginal hazard functions. In this case, tests of fit reduce to the fully parametric case.

5.4 Bivariate Sequential Data

Sequentially ordered survival times are of interest in many studies. For example, sequences of survival times may be the times between successive recurrent events such as bone fractures in cancer patients or pulmonary infections in persons with cystic fibrosis (Cook and Lawless, 2007, chapters 4, 6), the times between repeat admissions to a psychiatric facility (Kessing et al., 1998) or the duration of time spent in disease-free and subsequent relapse states for cancer patients (Lin et al., 1999; Cook et al., 2003). Problems arise with nonparametric estimation for sequential data when the survival times are not independent. This leads to dependent censoring and non-identifiability of the marginal distributions of the second and subsequent survival times (Lin et al., 1999; Schaubel and Cai, 2004a; Cook and Lawless, 2007). Another issue is the fact that in studies such as the colon cancer example in Section 4.6, a significant proportion of subjects may never have the first event. A similar pattern occurs in a nonrandomized clinical trial of adjuvant chemotherapy for breast cancer conducted by the International Breast Cancer Study Group (IBCSG). This study investigated the effectiveness of short duration (one month) versus long duration (six or seven months) chemotherapy (The Ludwing Breast Cancer Study Group, 1988). Cook et al. (2003) considered the times spent in remission and from relapse to death in the two treatment groups. In both examples, some individuals do not experience relapse and, indeed, may be cured of their disease. Thus, modeling of the distribution of the time to first event, T_1, and the joint distribution of T_1 and the time between the first and second events, T_2 should reflect this feature.

We proposed modeling and semiparametric estimation methods which overcome the difficulties caused by sequential data. A copula function is used to model the conditional probability $Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 < \infty)$ due to the fact that the second event cannot be seen if the subject never has the first event. A semiparametric estimation procedure is used to fit the copula function where the marginal distributions are left nonparametric. This also provides a means of checking parametric models for T_2.

Another possible approach to model bivariate sequential data is using a copula model for the truncated distribution $Pr(T_1 \leq t_1, T_2 \leq t_2 | T_1 \leq Q)$ where $Q < C_{\max}$ is some selected value. Hence, the model given in (4.26) was introduced and its relationship with (4.3) and (4.7) were developed.

5.4.1 Semiparametric Estimation

We proposed a new approach to the estimation of the joint distribution of sequentially observed survival times by considering copula models in which the marginal distributions are treated nonparametrically. This allows estimation of the marginal distributions of second or subsequent survival times and of the association among survival times for an individual. The presentation here focused on the case of two times, but the extension to $K \geq 3$ times $T_1, , ..., T_K$ is in principle straightforward. In this case, however, optimization of the log-likelihood function to obtain estimates of copula parameters α and parameter vectors λ_k^* $(k = 1, ..., K)$ for each marginal distribution may be challenging. A compromise procedure is to use the two-stage method recursively. Once $F_1(t_1)$ and $F_2(t_2)$ are estimated by the current two-stage approach (ignoring any later survival times), we can estimate $F_3(t_3)$ and α by pseudolikelihood with F_1, F_2 fixed and so on. Properties of this ad hoc approach would need to be investigated. For larger K and for larger sample size n, we recommend the use of piecewise-constant or spline-based hazard functions in copula models (He and Lawless, 2003). They have the advantage of flexibility for marginal distributions with a moderate number of parameters.

For the case where $K = 2$ the approach is readily extended to handle covariates through the adoption of Cox models for T_1, T_2. Semiparametric maximum likelihood estimates for baseline hazard functions (represented by λ_1^*, λ_2^*), regression parameters β_1, β_2 and copula parameters α can be obtained either by two-stage estimation or by simultaneous maximization of the likelihood for $\lambda_1^*, \lambda_2^*, \beta_1, \beta_2, \alpha$.

As for the semiparametric estimators based on parallel data, it remains a difficult and challenging problem to develop asymptotic theory for the estimators based on sequential lifetimes.

We also introduced a two-stage semiparametric estimation approach to fit the copula model C_α^Q for the truncated distribution given in (4.26) or (4.27), that is inspired by Shih

and Louis (1995). In the first stage, $F_{1Q}(t_1) = Pr(T_1 \leq t_1 | T_1 \leq Q)$ and $F_{2Q}(t_2) = Pr(T_2 \leq t_2 | T_1 \leq Q)$ are estimated nonparametrically, and in the second stage, the semiparametric estimate of the vector of dependence parameters α is obtained. In a future work, properties of this estimation method will be investigated.

5.4.2 Model Checking and Tests of Fit for Copula Models

Informal model checking for the copula family can be performed by comparing plots of the semiparametric and nonparametric fits of the conditional probability $Pr(T_2 > t_2 | T_1 \leq t_1)$ for $t_1 + t_2 \leq C_{\max}$ as we did in Section 4.6.

We can also carry out a parametric likelihood ratio test after embedding the proposed copula model in an expanded parametric family of copulas as we have done for parallel clustered data. However, in this case, it is not as easy to check parametric specifications for the distributions of T_2, so it is useful to use a semiparametric likelihood ratio or semiparametric pseudolikelihood ratio test. When the proposed and the expanded copula models are estimated by the semiparametric maximum likelihood estimation procedure described in Section 4.2.2, a semiparametric likelihood ratio test statistic can be used. When the models are estimated by two-stage semiparametric estimation procedure described in Section 4.2.1, the corresponding test is a semiparametric pseudolikelihood ratio test.

Work on semiparametric maximum likelihood (e.g. Murphy and van der Vaart, 2000) shows that in many settings likelihood ratio and pseudolikelihood ratio (Liang and Self, 1996) statistics for finite-dimensional parameters have chi-squared asymptotics. Simulations suggest that in the present setting, profile likelihood or pseudolikelihood ratio statistics for α parameters have distributions close to those of linear combinations of $\chi^2_{(1)}$ variables when sample size is large. A practical complication is that values on the boundary of the parameter space for α are often of interest, as illustrated in Section 4.6, where both the Clayton and the Gumbel-Hougaard copulas correspond to boundary points in a two-parameter copula family. In addition, even if limiting distributions correspond to linear combinations of $\chi^2_{(1)}$ variables, the distribution may depend on unknown parameter values and thus have to be estimated. For practical purposes we therefore recommend bootstrap simulation for the provision of p-values, variance estimates and confidence limits.

APPENDICES

Appendix A

Estimating Functions

Let θ be a $p \times 1$ parameter vector and $U(\theta) = \sum_{i=1}^{n} U_i(\theta)$ be a set of estimating functions $U_j(\theta) = \sum_{i=1}^{n} U_{ji}(\theta)$ $(j = 1, ..., p)$ that are functions of the observed data and θ. The aim is to obtain an estimate $\tilde{\theta}$ by solving the estimating equations $U(\theta) = 0$.

Assume $U(\theta)$ is an unbiased estimating function, i.e. $E[U(\theta)] = 0$ for all θ. Then, under some regularity conditions

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d MVN(0, C(\theta)) \tag{A.1}$$

where 0 is a zero vector,

$$C(\theta) = A(\theta)^{-1} B(\theta)(A(\theta)^{-1})^t, \tag{A.2}$$

$A_n(\theta) = -\frac{1}{n}\frac{\partial U(\theta)}{\partial \theta^t}$, $B_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U_i(\theta)U_i^t(\theta)$, $A(\theta) = \lim_{n\to\infty} E[A_n(\theta)]$ and $B(\theta) = \lim_{n\to\infty} E[B_n(\theta)] = \lim_{n\to\infty} \frac{1}{n} Var[U(\theta)]$.

The covariance matrix $C(\theta)$ can be estimated consistently by

$$C_n(\tilde{\theta}) = A_n(\tilde{\theta})^{-1} B_n(\tilde{\theta})(A_n(\tilde{\theta})^{-1})^t. \tag{A.3}$$

White (1982) considered the asymptotic properties of $\tilde{\theta}$ when the model on which the estimating equations are based is misspecified. Suppose the true distribution of the i.i.d. random variables is G. If there is a unique vector θ^* such that $E_G[U_i(\theta^*)] = 0$, under some regularity conditions

$$\sqrt{n}(\hat{\theta} - \theta^*) \rightarrow_d MVN(0, C(\theta^*)) \tag{A.4}$$

where $\tilde{\theta}$ is the solution to $U(\theta) = 0$, $C(\theta^*)$ is obtained when θ in (A.2) is replaced by θ^* and $A(\theta^*)$ and $B(\theta^*)$ are obtained when the expectations are taken with respect to G. Again, the covariance matrix $C(\theta^*)$ can be consistently estimated by $C_n(\tilde{\theta})$ given in (A.3).

References

[1] Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* 22 , 1299-1327.

[2] Andersen, P. K.; Ekstrom, C. T.; Klein, J. P.; Shu, Y. and Zhang, M.-J. (2005). A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biometrical Journal* 47, 815-824.

[3] Breymann, W.; Dias, A. and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance* 3, 1-14.

[4] Burke, M. D. (1988). Estimation of a bivariate distribution function under random censorship. *Biometrika* 75, 379-382.

[5] Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* 68, 417-422.

[6] Campbell, G. and Földes, A. (1982). Large sample properties of nonparametric bivariate estimators with censored data. In *Nonparametric Statistical Inference, Colloquia Methemetica-Societatis, János Bolyai,* eds. B. V. Gnedenko, M. L. Puri, and I. Vincze, Amsterdam: North-Holland, 103-122.

[7] Chen, X. and Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics* 33, 389-414.

[8] Chen, X. and Fan, Y. (2007). A model selection test for bivariate failure-time data. *Econometric Theory* 23, 414-439.

[9] Chen, X.; Fan, Y. and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* 101, 1228-1240.

[10] Chen, S. X. and Huang, T.-M. (2007). Nonparametric estimation of copula functions for dependence modelling. *The Canadian Journal of Statistics* 35, 265-282.

[11] Clayton, D. G. (1978). A model for association in bivatriate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65, 141-151.

[12] Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A* 148, 82-117.

[13] Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events.* Springer, New York.

[14] Cook, R. J.; Lawless, J. F. and Lee, K. A. (2003). Cumulative processes related to even histories. *Statistics and Operations Research Transactions* 27, 13-30.

[15] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

[16] Dabrowska, D. M. (1988). Kaplan-Meier estimates on the plane. *The Annals of Statistics* 16, 1475-1489.

[17] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. Académie Royale de Belgique. *Bulletin de la Classe des Sciences*, 5e Série 65, 274292.

[18] Denuit, M.; Purcaru, O. and van Keilegom, I. (2004). Bivariate Archimedean copula modelling for loss-ALAE data in non-life insurance. Working paper, Université catholique de Louvain.

[19] Dobrić, J. and Schmid, F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation. *Computational Statistics and Data Analysis* 51, 4633-4642.

[20] Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* 76, 312-319.

[21] Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis* 95, 119-152.

[22] Fermanian, J.-D.; Radulovic, D. and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* 10, 847-860.

[23] Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematical* 19, 194-226.

[24] Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sc* 4, 53-84.

[25] Fréchet, M. (1958). Remarques au sujet de la note précédente. *Comptes Rendus de l'Académie des Sciences de Paris Série I Math.* 246, 2719-2720.

[26] Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* 2, 1-25.

[27] Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* 85, 1-11.

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

[29] Genest, C.; Ghoudi, K. and Rivest, L.-P. (1998). Comment on a paper by E. W. Frees and E. A. Valdez entitled "Understanding relationships using copulas". *North American Actuarial Journal* 2, 143-149.

[30] Genest, C. and MacKay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionelles dont les marges sont données. *The Canadian Journal of Statistics* 14, 145-159.

[31] Genest, C.; Quessy, J.-F. and Rémillard, B. (2006a). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics* 33, 337-366.

[32] Genest, C.; Quessy, J.-F. and Rémillard, B. (2006b). On the joint asymptotic behaviour of two rank-based estimators of the association parameter in the gamma frailty model. *Statististics and Probability Letters* 76, 10-18.

[33] Genest, C.; Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré: Probabilités et statistiques* 44, 1096-1127.

[34] Genest, C.; Rémillard, B. and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: a review and a power study. *Insurance: Mathematics and Economics* 44, 199-213.

[35] Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel's family of extreme value distributions. *Statistics and Probability Letters* 8, 207-211.

[36] Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* 88, 1034-1043.

[37] Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transformation. *Statistics and Probability Letters* 53, 391-399.

[38] Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data. *The Canadian Journal of Statistics* 30, 557-571.

[39] Georges, P.; Lamy, A-G.; Nicolas, E.; Quibel, G. and Roncalli, T. (2001). Multivariate survival modeling: a unified approach with copulas. *Working paper, Crédit Lyonnais.*

[40] Ghoudi, K.; Khoudraji, A. and Rivest, L.-P. (1998). Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *The Canadian Journal of Statistics* 26, 187-197.

[41] Gijbels, I. and Mielniczuk, J. (1990). Estimating the density of a copula function. *Communications in Statistics - Theory and Methods* 19, 445-464.

[42] Glidden, D. V. (2000). A two-stage estimator of the dependence parameter for the Clayton-Oakes model. *Lifetime Data Analysis* 6, 141-156.

[43] Glidden, D. V. and Self, S. G. (1999). Semiparametric likelihood estimation in the Clayton-Oakes failure time model. *Scandinavian Journal of Statistics* 26, 363-372.

[44] Gumbel, E. J. (1960). Distribution des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de statistique de l'Université de Paris* 9, 171-173.

[45] He, W. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics* 59, 837-848.

[46] He, W. and Lawless, J. F. (2005). Bivariate location-scale models for regression analysis, with applications to lifetime data. *Journal of the Royal Statistical Society B* 67, 63-78.

[47] Hougaard, P. (1989). Fitting a multivariate failure time distribution. *IEEE Transactions and Reliability* 38, 444-448.

[48] Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer-Verlag, New York.

[49] Hougaard, P.; Harvald, B. and Holm, N. V. (1992). Measuring the similarities between lifetimes of adult Danish twins born between 1881-1930. *Journal of the American Statistical Association* 87, 17-24.

[50] Hsieh, J.-J., Wang, W. and Ding, A. A. (2008).Regression analysis based on semicompeting risks data. *Journal of the Royal Statistical Society B* 70, 3-20.

[51] Hsu, L. and Gorfine, M. (2006). Multivariate survival analysis for case-control family data. *Biostatistics* 7, 387-398.

[52] Huster, W. J., Brookmeyer, R. and Self, S. G. (1989). Modeling paired survival data with covariates. *Biometrics* 45, 145-156.

[53] Ibrahim, J. G.; Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer Verlag, New York, NY.

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

[55] Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* 94, 401-419.

[56] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd edition. John Wiley & Sons, New York.

[57] Kessing, L. V.; Olsen, E. W. and Andersen, P. K. (1998). Recurrence in affective disorder - anaysis with frailty models. *American Journal of Epidemiology* 149, 404-411.

[58] Klugman, S. A. and Parsa, R. (1995). Fitting bivariate loss distributions with Plackett's model. *Casualty Actuarial Society Ratemaking Seminar.*

[59] Klugman, S. A. and Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics* 24, 139-148.

[60] Kotz, S.; Johnson, N. L. and Balakrishnan, N. (2000). *Continuous Multivariate Distributions* 2nd edition. Wiley, New York.

[61] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* 2nd edition. John Wiley & Sons, Hoboken.

[62] Lawless, J. F. and Fong, D. Y. T. (1999). State duration models in clinical and observational studies. *Statistics in Medicine* 18, 2365-2376.

[63] Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* 37, 1137-1153.

[64] Li, Y.; Prentice, R. L. and Lin, X. (2008). Semiparametric maximum likelihood estimation in normal transformation models for bivariate survival data. *Biometrika* 95, 947-960.

[65] Liang, K.-Y. and Self, S. G. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society B* 58, 785-796.

[66] Lin, D. Y.; Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika* 86, 59-70.

[67] Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* 80, 573-581.

[68] Martinussen, T. and Pipper, C. B. (2005). Estimation in the positive stable frailty Cox proportional hazards model. *Lifetime Data Analysis* 11, 99-115.

[69] Moertel, C. G.; Fleming, T. R. and McDonald, J. S. (1990). Levamisole and Fluorouracil for adjuvant therapy of restriced colon carcinoma. *New England Journal of Medicine* 322, 352-358.

[70] Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* 95, 449-465.

[71] Nelsen, R. B. (2006). *An Introduction to Copulas.* 2nd edition. Springer, New York.

[72] Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society B* 44, 414-422.

[73] Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* 73, 353-361.

[74] Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84, 487-493.

[75] Oakes, D. (1994). Multivariate survival distributions. *Journal of Nonparametric Statistics* 3, 343-354.

[76] Oakes, D. (2005). On the preservation of copula structure under truncation. *The Canadian Journal of Statistics* 33, 465-468.

[77] Pena, E. (1998). Smooth goodness-of-fit tests for composite hypothesis in hazard-based models. *The Annals of Statistics* 26, 1935-1971.

[78] Pipper, C. B. and Martinussen, T. (2003). A likelihood based estimating equation for the Clayton-Oakes model with marginal proportional hazards. *Scandinavian Journal of Statistics* 30, 509-521.

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

[80] Prentice, R. L., Zoe Moodie, F. and Wu, J. (2004). Hazard-based nonparametric survivor function estimation. *Journal of the Royal Statistical Society B* 66, 305-319.

[81] Pruitt, R. G. (1990). Strong consistency of self consistent estimators: general theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, School of Statistics.

[82] Pruitt, R. G. (1991). On negative mass assigned by the bivariate Kaplan-Meier estimator. *The Annals of Statistics* 9, 879-885.

[83] Reid, N. (1981). Estimating the median survival time. *Biometrika* 68, 601-608.

[84] Romeo, J. S.; Tanaka, N. I. and Pedroso-de-Lima, A. C. (2006). Bivariate survival modeling: a Bayesian approach based on copulas. *Lifetime Data Analysis* 12, 205-222.

[85] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* 23, 470-472.

[86] Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *The Annals of Statistics* 4, 912-923.

[87] Sahu, S. K. and Dey, D. K. (2000). A comparison of frailty and other models for bivariate survival data. *Lifetime Data Analysis* 6, 207-228.

[88] Schaubel, D. E. and Cai, J. (2004a). Nonparametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine* 23, 1885-1900.

[89] Schaubel, D. E. and Cai, J. (2004b). Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika* 91, 291-303.

[90] Self, G. S. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82, 605-610.

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

[92] Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 51, 1384-1399.

[93] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris* 8, 229-231.

[94] Sungur, E. A. (2002). Some results on truncation dependence invariant class of copulas. *Communications in Statistics - Theory and Methods* 31, 1399-1422.

[95] The Ludwing Breast Cancer Study Group (1988). Combination adjuvant chemotherapy for hode-positive breast cancer: inadequacy of a single perioperative cycle. *New England Journal of Medicine* 319, 677-683.

[96] Tsai, W. Y.; Leurgans, S. and Crowley, J. J. (1986). Nonparametric estimation of a bivariate survival function in presence of censoring. *The Annals of Statistics* 14, 1351-1365.

[97] van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics* 24, 596-627.

[98] van der Laan, M. J.; Hubbard, A. E. and Robins, J. M. (2002). Locally efficient estimation of a multivariate survivor function in longitudinal studies. *Journal of the American Statistical Association* 97, 494-507.

[99] Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS. *Biometrika* 83, 507-518.

[100] Wang, M. C. and Chang, S. H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association* 94, 146-153.

[101] Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika* 87, 879-893.

[102] Wang, W. and Wells, M. T. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika* 84, 863-880.

[103] Wang, W. and Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85, 561-572.

[104] Wang, W. and Wells, M. T. (2000a). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* 95, 62-72.

[105] Wang, W. and Wells, M. T. (2000b). Estimation of Kendall's tau under censoring. *Statistica Sinica* 10, 1199-1215.

[106] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1-25.