FDZ Methodenreport

Methodological aspects of labour market data

No. 2/2009(EN)

A copula model for dependent competing risks

Simon M. S. Lo and Ralf A. Wilke

A copula model for dependent competing risks.*

Simon M. S. Lo†

Ralf A. Wilke‡

February 2009

Abstract

Many popular estimators for duration models require independent competing risks or independent censoring. In contrast, copula based estimators are also consistent in presence of dependent competing risks. In this paper we suggest a computationally convenient extension of the Copula Graphic Estimator (Zheng and Klein, 1995) to a model with more than two dependent competing risks. We analyse the applicability of this estimator by means of simulations and real world unemployment duration data from Germany. We obtain evidence that our estimator yields nice results if the dependence structure is known and that it is a powerful tool for the assessment of the relevance of (in-)dependence assumptions in applied duration research.

Keywords: Archimedean copula, dependent censoring, unemployment duration

JEL: C41, C51

*We thank Simon Lee for contributing several important ideas to our work and Melanie Arntz and Anton Nivorozhkin for helpful comments. Wilke is supported by the Economic and Social Research Council through the grant *Bounds for Competing Risks Duration Models using Administrative Unemployment Duration Data* (RES-061-25-0059). This work uses the IAB Employment Subsample (IABS 2001-R01) of the Research Data Centre at the Institute of Employment Research (IAB). The IAB does not take any responsibility for the use of its data.

†University of Freiburg, E–mail: losimonms@yahoo.com.hk

‡University of Nottingham, E–mail: ralf.wilke@nottingham.ac.uk

1 Introduction

Applied economic research usually faces the challenge to model an empirical problem in such a way that it is not too complex but still realistic. With regard to duration analysis, the complexity of the underlying problem often requires a competing risks structure. As an example, we may want to study the effect of unemployment compensation transfers on the duration of unemployment. The model would be too narrow if it focuses on transitions to employment only because the policy can have multiple effects. There may be also impacts on the transitions to other risks such as the timing of early retirement or assignment into active labour market programmes. A multivariate competing risks model is in this case more appropriate for the empirical analysis. Unfortunately, observed data alone is not sufficient to identify the marginal distributions of the latent variables if the dependence structure between risks is unknown. This well known fundamental identification problem cannot be resolved (Cox, 1962, Tsiatis, 1975). If one is not willing to impose identifying assumptions, it is only possible to obtain bounds for the marginal distributions (Peterson, 1976). Parametric or semiparametric versions of the proportional hazard (PH) model or the mixed proportional hazard model (MPH) are popular in applied economic research. These approaches have well explored properties and they are rather convenient to apply. These models are identified if among other things the covariate structure possesses certain properties (Heckman and Honoré, 1989, Abbring and van den Berg, 2003). By using these models one imposes implicit assumptions on the marginal distributions of the latent variables and their dependence structure. For instance, the factor-loading specifications are often used to estimate the MPH model. See Van den Berg (2001) for a detailed discussion. Canals-Cerdá and Gurmu (2007) propose a rather different estimation technique to approximate the dependence structure of a frailty model in a nonparametric way. Another popular approach in applied work is to assume independence of latent variables. In this case, the famous product limit estimator (Kaplan and Meier, 1958) is consistent. The popularity of this estimator certainly stems to a large extend from the fact that it does not require strong assumptions on the marginal distributions of the latent variables. Alternatively, if one wants to avoid the independence assumption one can also model the joint dependence structure by means of a copula function.

Copula based models represent a wide model class and one can show that popular duration models are in fact special cases of the copula model. As there are many different families of copulas (Nelsen, 2006), the model allows for flexible specification of the dependence structure between competing random variables. Identification and estimation are already analysed in several contributions. Zheng and Klein (1995) prove identification of the marginal distributions for a model with two risks and a known copula function with known parameters. Their nonparametric estimator is known as the Copula Graphic Estimator. Carrière (1995) proves identification in

presence of more than two risks. Although, his nonparametric copula approach is valid under fairly mild conditions, it has an important practical limitation. By solving a system of simultaneous differential equations, the computation time increases substantially with the number of competing risks. The need for substantial computational resources to implement multivariate copula models is a general difficulty and not specific to duration models (Zimmer and Klein, 2006). Rivest and Wells (2001) suggest a martingale approach under the additional assumption that the copula is Archimedean and that failure times are distinct for all observations. They derive a closed form solution and thus the implementation is rather convenient. Unfortunately, their implementation does not work in applications with non distinct observations. The same applies to an extension of this estimator suggested by Braekers and Veraverbeke (2006).

Although copula functions are recently more frequently used in Statistics and Finance (Mikosch, 2006), research about copula based duration models is mainly driven by new developments in biometrics and mathematical statistics. We are not aware of an application to duration data in economics or social sciences alike. This paper therefore also shows the benefits of copulas for applied economic duration analysis. As an example we choose unemployment duration because of good data availability and a large potential user group in economics and social sciences. Our application to estimate the effect of an unemployment insurance reform aims at providing new insights to the reader whether the use of copulas contributes to a broader assessment of policy reform effects.

We see the following contributions in this paper:

- We extent the Copula Graphic estimator (Zheng and Klein, 1995) to a model with more than two competing risks when the copula is Archimedean. Our implementation works with common data structures as it does not require non distinct observations. We reason that our risk pooling method is computationally convenient. Moreover, we show that the estimator is consistent.

- We demonstrate the applicability and nice finite sample properties of our estimator with the help of simulations.

- We apply our framework by estimating the effect of a reduction in unemployment benefit entitlements on the duration of unemployment in Germany using a model with three competing risks. Our results suggest that the magnitude of the estimated reform effect is sensitive to the assumed dependence structure while the estimated sign is often insensitive.

The paper is structured as follows. Section 2 presents our estimation framework. Section 3 presents results of the simulation study, followed in the next section by an empirical illustration. The last section contains some conclusions.

2 Model

2.1 Framework

Let $(T_1, \ldots, T_J) \in \mathbb{R}^{\mathrm{J}}_{++}$ be latent duration times of risk $j = 1, \ldots, J$ in a J-dimensional competing risks model. We can only observe T_j if $T_j < T_i$ for all $i \neq j$. All other T_i are not observed. $(T_1, \ldots, T_J)$ could depend on each other. We assume that T_j is an unknown continuous function of X and U_j:

$$T_j = \psi_j(X, U_j). \tag{1}$$

X is a k-dimensional vector of observable variables. U_j is an unobservable variable and is usually called unobserved heterogeneity. X and U_j are independent. $U_1, \ldots, U_J$ can be dependent. If $U_i \equiv U_j$ for all $i \neq j$, it implies a correlation of $+1$ between all the unobservables and the conditional joint distribution of durations is therefore degenerate.

Let $S_j : R_{++} \to [0, 1]$ be the unknown continuous and strictly decreasing marginal survival function of T_j:

$$S_j(t_j) = \Pr(T_j > t_j) = r_j. \tag{2}$$

$r_j \in R_j$ is defined as the relative position or rank order of $t_j \in T_j$. $R_j = S_j(\psi(X, U_j))$ is therefore a uniformly distributed variable in $[0, 1]$. Conditional on $X = x$, $S_j(t_j|x)$ is also continuous and strictly decreasing. The conditional rank of T_j is then

$$\begin{aligned} R_{j|x} &= S_j(\psi_j(x, U_j)) = (S_j \circ \psi_j)(x, U_j) \\ &= G_j(U_j), \end{aligned}$$

where $G_j : \mathbb{R} \to [0, 1]$ is a continuous function which is independent of X. As a special case, if ψ_j is monotone in U_j then G_j is the survival function of U_j and the conditional rank of T_j is determined by the rank of U_j only. In the following, we focus our discussion on the case without conditioning on X, unless it is necessary.

We assume that the basic dependence structure of T_j is generated by a known copula, which is a joint distribution of the ranks of the duration variables. The J-copula, $C^J : [0, 1]^J \to [0, 1]$ is defined by

$$C^J(r_1, \ldots, r_J) = \Pr(R_1 \leq r_1, \ldots, R_J \leq r_J). \tag{3}$$

The copula relates only the ranks of different duration variables. Specification of the copula does not require a known functional form for the marginal survival function. The copula determines

therefore the basic dependence structure between the variables T_j. Note that this does not rely on the marginal distribution of the risks. Conditioning on $X = x$, the copula is denoted as C_x^J.

If the marginal survival functions are given by (2), the joint survival function, $S(t_1, \ldots, t_J) = \Pr(T_1 > t_1, \ldots, T_J > t_J)$, is uniquely determined by substituting (2) into (3)

$$\begin{aligned} C^J(S_1(t_1), \ldots, S_J(t_J)) &= \Pr(S_1(T_1) \leq S_1(t_1), \ldots, S_1(T_J) \leq S_J(t_J)) \\ &= S(t_1, \ldots, t_J). \end{aligned} \tag{4}$$

Equivalently, given any $S(t_1, \ldots, t_J)$ and $S_j(t_j)$, there is a unique C^J such that (4) holds. Uniqueness is proved by Sklar's theorem (Schweizer and Sklar, 1983).

Given the copula and the marginal survival functions, the joint survival function $S(t) = S(t, \cdots, t)$ and the cause specific cumulative incidence curve (CIC), $Q_j(t_j)$ are given by

$$\begin{aligned} S(t) &= P(T_1 > t, \ldots, T_J > t) \\ &= \int_0^{r_J(t)} \cdots \int_0^{r_1(t)} dC^J(r_1, \ldots, r_J) \end{aligned} \tag{5}$$

$$\begin{aligned} Q_j(t_j) &= P(T_j \leq t_j, T_j < \min_{i \neq j}\{T_i\}) \\ &= \int_0^{\zeta_J(r_j)} \cdots \int_0^{\zeta_{j+1}(r_j)} \int_{r_j(t_j)}^{1} \int_0^{\zeta_{j-1}(r_j)} \cdots \int_0^{\zeta_1(r_j)} dC^J(r_1, \ldots, r_J), \end{aligned} \tag{6}$$

where $\zeta_k(r_j) = S_k(S_j^{-1}(r_j))$ for all $k \neq j$. Note that the inverse exists since S_j is continuous and strictly decreasing.

In an application, we face the reversed problem as usually estimates for $S(t)$ and $Q_j(t_j)$ for all j are available only. Our aim is then to determine the unknown marginal survival functions $\{S_1(t_1), \ldots, S_J(t_J)\}$ using $\{S(t), Q_1(t_1), \ldots, Q_J(t_J), C^J\}$. While $S(t)$ and $Q_j(t_j)$ can be estimated, the true copula function C^J has to be known or to be assumed. There are many different classes of copulas describing different basic dependence structures of the variables. Nelsen (2006) provides a comprehensive overview over different families. Copulas can differ in their functional form and in their parameter(s). Both determine the dependence degree between the T_j's. One can also show that there is a direct link between the copula model under additional assumptions and the popular duration models of applied economic research. We now present several important copula functions.

If T_i is independent of T_j for all $i \neq j$, any copula reduces to the product copula

$$C^J(r_1, \ldots, r_J) = \prod_{j=1}^{J} G_j(U_j). \tag{7}$$

The Kaplan-Meier estimator requires independence.

The Archimedean copula is defined by

$$C^J(r_1, \ldots, r_J) = \phi^{-1}(\phi(r_1) + \ldots + \phi(r_J)), \tag{8}$$

where $\phi(r) : [0,1] \rightarrow \mathbb{R}_+$ is the so called copula generator, a strictly decreasing and twice differentiable continuous function with $\phi(1) = 0$. The Archimedean class is important as it covers a wide range of families. Moreover, Archimedean copulas are easy to construct and have nice properties as they are symmetric (i.e. $C^2(r_1, r_2) = C^2(r_2, r_1)$ for $J = 2$) and they are associative (i.e. $C^2(C^2(r_1, r_2), r_3) = C^2(r_1, C^2(r_2, r_3))$ for $J = 2$). As a result C^J can be constructed step by step from a 2-copula by

$$\begin{aligned} C^J(r_1, \ldots, r_J) &= C^2(C^{J-1}(r_1, \ldots, r_{J-1}), r_J) \\ &= C^2(C^2(C^{J-2}(r_1, \ldots, r_{J-2}), r_{J-1}), r_J) \\ &= \ldots \\ &= C^2(C^2(...C^2(C^2(r_1, r_2), r_3), \ldots), r_J). \end{aligned}$$

Schweizer and Sklar (1983) denote this procedure as serial iteration of the Archimedean 2-copula which implies that the dependence structure between all r_j is the same. When we condition on $X = x$, (8) becomes

$$C_x^J(r_1, \ldots, r_J|x) = \phi^{-1}(\phi(G_1(u_1)) + \ldots + \phi(G_J(u_J))). \tag{9}$$

The Archimedean class has many different sub-classes and families. One special subclass is the frailty model (Oakes, 1989). Suppose that the following conditions hold:

1. $\psi_j(X, U_j) = \tilde{\psi}_j(X)U_j$, such that U_j is monotone in ψ_j;
2. All T_j have an exponential marginal survival distribution as $S_j(t_j|x) = \exp[\Lambda_j(t_j)\tilde{\psi}_j(x)u_j]$ where $\Lambda_j(t_j)$ is the integrated baseline hazard function;
3. The copula generator from an Archimedean copula is a Laplace transformation of a joint distribution function of the U_js, denoted by $G(u)$ with $u = [u_1, \ldots, u_J]$.

Then, the joint survival function is a mixed proportional hazard model of the form

$$\begin{aligned} S(t, \ldots, t|x) &= C^J(r_1, \ldots, r_J|x) \\ &= \int \exp[\Lambda_1(t)\tilde{\psi}_1(x)u_1 + \ldots + \Lambda_J(t)\tilde{\psi}_J(x)u_J]dG(U). \end{aligned} \tag{10}$$

One important subfamily of the Laplace transformation is the Clayton copula (Clayton, 1978). In this case $G(u)$ is a gamma distribution, $U \sim \Gamma(1/\theta, 1)$ and the copula generator is $\phi(s) =$

$s^{-1/\theta} - 1$, with $\theta > 0$. See also Table 1 which lists several common bivariate Archimedean copulas with a single parameter θ. For more copulas see Nelsen (2006). The Frank copula is often used in applied work because of its capability to incorporate all possible degrees of dependence.

Table 1: One parameter families of Archimedean copulas [1]

Family	$\tilde{C}(u,v)$	$\phi^{-1}(s)$	Laplace Transform
Clayton	$\left[\max(u^{-\theta} + v^{-\theta} - 1, 0)\right]^{-1/\theta}$	$(\frac{1}{1+\theta s})^{1/\theta}$	yes
Ali-Mikhail-Haq	$\frac{uv}{1-\theta}(1-u)(1-v)$	$\frac{1-\theta}{e^s-\theta}$	yes
Frank	$-\frac{1}{\theta}\ln(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1})$	$-\frac{1}{\theta}\ln(1 + e^{-s}(e^{-\theta} - 1))$	yes
Unknown	$\max(1 + \theta/\ln[e^{\theta/(u-1)} + e^{\theta/(v-1)}], 0)$	$\frac{1}{\theta}\ln(s) + 1$	no

[1] $\tilde{C}(u,v)$ is defined here as $\tilde{C}(H_1(t_1), H_2(t_2))$, with $H_j(t_j)$ as the marginal distribution. The copula C can be obtained from $\tilde{C}$ by the equation $C(S_1(t_1), S_2(t_2)) = S_1(t_1) + S_2(t_2) - 1 + \tilde{C}(1 - S_1(t_1), 1 - S_2(t_2))$.

2.2 Identification and Estimation

Zheng and Klein (1995) and Carrière (1995) prove identification of the marginal survival functions $S_j(t_j)$ if the copula is known. While Carrière's model can have several competing risks, Zheng and Klein's proof applies to the case of two risks only. Given equation (6), Carrière's approach is based on the fact that the derivative of the conditional marginal survival functions can be identified by solving a system of J nonlinear differential equations

$$\frac{d}{dt}Q_j(t) \quad = \quad \frac{d}{dt}S_j(t) \times \frac{\partial C^J(S_1, \ldots, S_J)}{\partial S_j}\bigg|_{t_1=\ldots=t_J=t}, \tag{11}$$

for $j = 1, \ldots, J$. Starting with the initial condition $S_j(0) = 1$, $S_j(t)$ can be recursively determined. However, the practical implementation of this approach can be rather difficult if there are several risks. For instance, let d be the number of numerical steps required to obtain $r_j(t)$ given $r_k(t)$ for $k \neq j$. Then we need d^J steps at each t to solve simultaneously for the J unknowns. This can be rather demanding if J is large. Moreover, his numerical algorithm determines $S_j(t + \Delta) - S_j(t)$ from $Q_j(t + \Delta) - Q_j(t)$ with $\Delta > 0$. Then $S_j(t + \Delta)$ is computed by adding the estimate for the difference to $S_j(t)$. This approximation is imprecise if the observed failures for each risks are not close to each other, i.e. Δ is not small. Moreover, the approximation error increases with t.

As a more practical solution, we suggest an extension of the implementation proposed by Zheng and Klein (1995). For the case $J = 2$, Zheng and Klein determine $S_j(t)$ by solving equations (5) and (6) directly. We suggest for the case $J > 2$, that $S_j(t)$ can be computed by pooling all other

risks $k \neq j$. The numerical algorithm proposed by Zheng and Klein for a two risks model can then be directly applied to determine $S_j(t)$. Repeating this pooling procedure we can compute all $S_j(t)$ separately. To solve for the J unknowns at each t, we need $J \times d^2$ steps only.

In order to keep things simple, we now illustrate our risk pooling method for the case $J = 3$. The idea is to pool the variables T_2 and T_3 to form a new variable $T_{23} = \min\{T_2, T_3\}$. The marginal survival function of T_{23} is defined as $S_{23}(t) = \Pr(T_{23} > t)$. If there is a survival copula between the variables T_1 and T_{23}:

$$C_1^2(S_1(t_1), S_{23}(t_{23})) = S(t), \tag{12}$$

we can directly apply Zheng and Klein's approach to compute $S_1(t_1)$ and $S_{23}(t_{23})$. This is done by solving the following two equations

$$\begin{aligned} S(t) &= \int_0^{r_{23}(t)} \int_0^{r_1(t)} dC_1^2(S_1, S_{23}) \\ Q_1(t) &= \int_0^{\zeta_{23}(r_1)} \int_{r_1(t)}^1 dC^2(S_1, S_{23}), \end{aligned}$$

where $r_{23}(t) = S_{23}(t)$ and $\zeta_{23}(r_1) = S_{23}(S_1^{-1}(r_1))$. The first equation holds because

$$\begin{aligned} \int_0^{r_{23}(t)} \int_0^{r_1(t)} dC_1^2(S_1, S_{23}) &= \Pr(S_1 \leq r_1(t), S_{23} \leq r_{23}(t)) \\ &= \Pr(T_1 > t, T_{23} > t) \\ &= \Pr(T_1 > t, T_2 > t, T_3 > t) = S(t) \end{aligned}$$

by noting that $T_{23} = \min\{T_2, T_3\}$. And thus we have $T_{23} > t$ if and only if $T_2 > t$ and $T_3 > t$. Similarly, the second equation holds because

$$\begin{aligned} \int_0^{\zeta_{23}(r_1)} \int_{r_1(t)}^1 dC_1^2(S_1, S_{23}) &= \Pr(S_1(t) > r_1(t), S_{23} \leq \zeta_{23}(r_1)) \\ &= \Pr(T_1 \leq t, T_{23} > T_1) \\ &= \Pr(T_1 \leq t, T_2 > T_1, T_3 > T_1) = Q_1(t). \end{aligned}$$

$S_2(t)$ and $S_3(t)$ can be obtained in a similar way by plugging in the relevant functions in equation (12). For this purpose we need the variables $T_{13} = \min(T_1, T_3)$ and $T_{12} = \min(T_1, T_2)$ and we need the copulas $C_2^2(S_2(t_2), S_{13}(t_{13}))$ and $C_3^2(S_3(t_3), S_{12}(t_{12}))$ respectively. This risk pooling method can be easily extended to the case of $J > 3$.

Unfortunately, a pooled 2-copula is generally inconsistent with the non-pooled 3-copula of T_1, T_2, T_3. This means the copula $C_1^2(S_1(t_1), S_{23}(t_{23}))$ in (12) may not exist (Genest et al., 1995). There are, however, necessary conditions such that (12) holds: For a known 3-copula $C^3(r_1, r_2, r_3)$, there exist all 2-copulas $C_{ij}(r_i, r_j) = S_{ij}(t)$ such that

$$C^3(r_1, r_2, r_3) = C_1^2(r_1, C_{23}(r_2, r_3)) = C_2^2(r_2, C_{13}(r_1, r_3)) = C_3^2(r_3, C_{12}(r_2, r_3)).$$

In this case C_{ij} is compatible with C^3 (Nelsen, 2006). While we are not aware of general conditions for the compatibility of copula functions, it is evident that any symmetric and associative copula is compatible. For this reason, the Archimedean class satisfies the required properties for the risk pooling method.

We can now carry over the identification strategy of Zheng and Klein (1995) to a model with more than two risks: the marginal distributions S_j for $j = 1, \ldots, J$ can be identified by the risk pooling method as outlined above, if the copula of the pooled variable is compatible. This is for example the case if

(**C1**) at least $J-2$ variables T_j, $j = 1, ..., J$, are independent and the copula between the two dependent variables is known; or

(**C2**) the J-copula is known and belongs to the Archimedean class.

Note that we cannot show identification of the Zheng and Klein approach for the general case in presence of more than two risks. Although Carrière (1995) proves identification for the general case, his implementation has disadvantages in an application as outlined above. We see our risk pooling approach as an interesting implementation for applied research as it is computationally more convenient and does not rely on specific requirements on the data structure. The model is convenient because it permits the focus on the estimation of the relevant risk. All the other risks can be pooled to decrease the computing time. The model therefore still allows for dependence on the other risks, but it assumes that this dependence structure has some regularities.

We finish this section by elaborating the estimation of the risk pooling approach, which is in fact a multivariate version of Zheng and Klein's Copula-Graphic Estimator. Suppose we have $i = 1, \ldots, n$ observations. The data generating process yields T_{ij} with $i = 1, ..., n$ and $j = 1, ..., J$. Due to the competing risks model only $\min_j\{T_{ij}\}$ for all i can be observed. Let S_n and Q_{jn} be estimators for $S(t) = P(T_1 > t, ..., T_J > t)$ and $Q_j(t) = P(T_j \leq t, T_j \leq T_1, \ldots, T_j \leq T_n)$ for $j = 1, ..., J$. Then the estimator for $S_j(\cdot)$ is the solution to

$$r_{jn}(t) \;=\; \operatorname{argmin}\left(A_n + B_{jn}\right), \tag{13}$$

where

$$\begin{aligned} A_n = &\left(S_n(t) - \int_0^{r_J(t)} \ldots \int_0^{r_1(t)} dC^J(r_1, \ldots, r_J)\right)^2 \\ B_{jn} = &\left(Q_{jn}(t) - \int_0^{\zeta_J(r_j)} \ldots \int_0^{\zeta_{j+1}(r_j)} \int_{r_j(t)}^1 \int_0^{\zeta_{j-1}(r_j)} \ldots \int_0^{\zeta_1(r_j)} dC^J(r_1, \ldots, r_J)\right)^2. \end{aligned}$$

$S_{jn}(\cdot)$ is therefore a function of $S_n(\cdot), Q_{1n}(\cdot), \ldots, Q_{Jn}(\cdot)$ and the copula C. Given that the estimates for the marginal distributions are the solution to well behaved objective functions, it is

straightforward that the consistency of Zheng and Klein's Copula-Graphic Estimator carries over to the multivariate case.

Corollary 1 *Let S_n and Q_{jn} be consistent estimators for S and Q_j. Then if one of the conditions **C1** or **C2** are met, S_{nj} as given by the solution to problem (13) is a consistent estimator for S_j for $j = 1, ..., J$.*

The proof of Corollary 1 can be based on Theorem 4.1.2 of Amemiya (1985) applied to objective function (13) provided that the estimators converge to nonstochastic functions and provided that the solution is unique. Moreover, the support of $r_j(t)$ is compact and the objective function is continuous. Similar tools could be applied to derive more asymptotic properties of the estimator.

2.3 Covariate Effects

Using the above framework we can also estimate the effect of a covariate change on the marginal distributions. For the purpose of illustration, we consider a treatment effect setting. We define a treatment dummy as $D = 1$ when an individual receives treatment and $D = 0$ otherwise. For simplicity we assume that the treatment is independent of all other observables and unobservables (X, U). Then, the conditional treatment effect, $\Delta_j(t|x)$, on $S_j(t|x)$ is simply

$$\Delta_j(t|x) \quad = \quad S_j(t|x, D = 1) - S_j(t|x, D = 0)$$

for $j = 1, \ldots, J$. $S_j(t|x, D = 1)$ and $S_j(t|x, D = 0)$ can be estimated using the above framework using the conditional copulas $C^J_{\{x,D=1\}}$ and $C^J_{\{x,D=0\}}$. Note that independence between X and U_j does not imply that the joint distribution of $(U_1, \ldots, U_J)$ is also independent of X. This follows from the fact that copula functions can depend on X. This implies that the latent variables T_i and T_j do not necessarily have the same dependence structure conditional to $D = 0$ and $D = 1$. However, applied research often ignores this possibility since the Kaplan-Meier estimator requires that the copula is independent of the covariates. The copula model is therefore compatible with empirical settings in which the treatment has not only an effect on the marginal distributions but also on the dependence structure. This includes the special case where there is just a change in the dependence structure without any change in the marginal distributions. In this case we observe a change in the joint survival function and in the CIC's. A correctly specified copula model would, however, identify that the marginal distributions are invariant. In contrast, the Kaplan-Meier estimator would suggest a change in the marginal distribution due to the treatment.

In the next section we explore how our suggested implementation of the copula based estimator performs in a simulation study. We pay special attention to the estimated treatment effect under different dependence structures and use a model with independent risks as a comparison benchmark.

3 Simulations

In order to investigate the finite sample properties of our risk pooling method, we simulate a model with three risks and an independent treatment dummy with Prob($D = 0$)=1-Prob($D = 1$)=0.5. The data is simulated by using a Frank copula and the copula parameter is chosen such that the correlation between the ranks of all risks is 0.5. Risk 1 is generated by a logistic distribution, risks 2 and 3 follow exponential distributions (see table 2). The parameters of the marginal distributions ($H_j(t)$) are chosen to produce different degrees of censoring. We repeat the simulations for three different sample sizes (50, 500, 1000) and we draw 500 independent samples for each sample size. Note that in this simulation design, the joint survival curve is 0.77, 0.51 and 0.09 at $t = 0.07, 0.24$ and 1.2, respectively. This means that at $t = 1.2$, there are only about 9% of the observations remaining in the risk set. In our simulation we want to estimate the six marginal distributions and the treatment effect for each risk.

Table 2: Simulation design.

	Distribution	Parameters	
		Control Group	Treatment Group
Risk 1	Logistic	(0.6, 1.4)	(0.9, 1.2)
Risk 2	Exponential	1.0	1.5
Risk 3	Exponential	0.8	0.8

As already discussed by Zheng and Klein (1995) the reliability of the Copula Graphic Estimator is affected by the degree of censoring in the data. In order to illustrate this further we report both the amount of censoring and the finite sample bias of the estimated marginal distributions in Table 3. The third and the fourth column of Table 3 present the degree of non-censoring for the three risks j which is defined by $\%Q_j = Q_j(t)/H_j(t)$. These numbers are reported for the treatment (D=0), the control group (D=1) and for three different durations (t=0.07, 0.24, and 1.20). It is apparent that the degree of censoring is not constant and note that $\sum_j \%Q_j(\infty) = 1$ as $H_j(\infty) = 1$ for all j. This also explains why the sum is not equal to one for shorter durations. In order to see how censoring and the properties of estimates are related, we compute the finite sample bias of the marginal survival functions defined by $B_n(S_{nj}(t)) = E_n(\hat{S}_j(t) - S_j(t))$ using the 500 estimates for the case of 500 observations. The bias is reported in columns 6 and 7 in Table 3. It generally increases with the share of censored observations.

Table 3: Degree of non-censoring and finite sample bias of the estimated marginal survival curves with 500 observations.

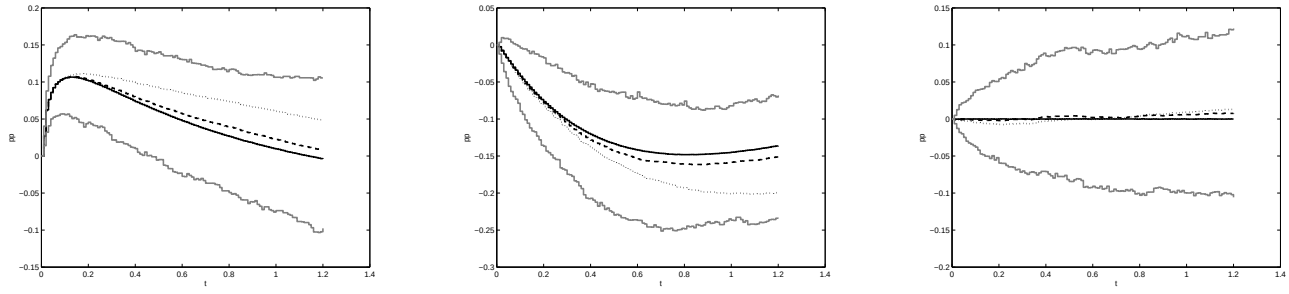
	Non-censored data	$D=0$	$D=1$	Finite sample bias	$D=0$	$D=1$
$t=0.07$	$\%Q_1$	0.92	0.89	$B_n(S_{n1}(t))$	0.0018	-0.0002
	$\%Q_2$	0.73	0.87	$B_n(S_{n2}(t))$	0.0010	-0.0003
	$\%Q_3$	0.72	0.81	$B_n(S_{n3}(t))$	0.0009	0.0015
$t=0.24$	$\%Q_1$	0.85	0.73	$B_n(S_{n1}(t))$	0.0023	-0.0025
	$\%Q_2$	0.57	0.72	$B_n(S_{n2}(t))$	0.0055	0.0036
	$\%Q_3$	0.54	0.63	$B_n(S_{n3}(t))$	0.0081	0.0001
$t=1.20$	$\%Q_1$	0.67	0.46	$B_n(S_{n1}(t))$	0.0118	0.0257
	$\%Q_2$	0.41	0.56	$B_n(S_{n2}(t))$	0.0224	0.0068
	$\%Q_3$	0.35	0.33	$B_n(S_{n3}(t))$	0.0347	0.0383

In applications we are often interested in the marginal effect of covariates. For this reason we compute the difference between the estimate for the marginal survival curve of the treatment group and the control group as an estimate for the treatment effect. Figure 1 shows the true treatment effect, the mean estimated treatment effect and the 5% and 95% quantiles of the distribution of estimated treatment effects for the three risks obtained from the 500 samples with 500 observations. It is apparent that the treatment effect varies across three risks and it is not constant with elapsed duration.

The figure shows that the correctly specified copula based estimator is close to the true values, however, there is a small bias in some cases. As a benchmark comparison, we also report the mean estimate of the treatment effect if we assumed independence of risks. In an application the results with assumed independence are very similar to the Kaplan-Meier (KM) estimator. For this reason we refer to this as the KM equivalent estimator in what follows. It is apparent from the figure that the KM equivalent estimator is more biased for risks 1 and 2 than the correctly specified estimator. As the reported quantiles of the distributions of estimates are wide, the figure also shows that the second moment of the distribution is by means not negligible. This is partly because the estimated treatment effect is a sum of two estimates. To get a better understanding of the finite sample properties, we compute the mean squared error (MSE) of the estimator for the treatment effect. Table 4 presents the MSE for different samples sizes and different durations. There is strong evidence that the MSE decreases with the sample size, which is mainly driven by the decrease in the variance. We also observe that the small systematic bias does not vanish at the same speed as the variance tends to zero. A similar finite sample bias is also observed by Zheng and Klein (1995) and it may be due to some numerical approximations in the implementation

Figure 1: Results of simulations with 500 observations: true treatment effect (solid line); mean estimated treatment effect (dashed line); 5% and 95% quantiles (grey lines); mean estimated treatment effects with assumed independence (dotted line).

(a) Risk 1 (b) Risk 2 (c) Risk 3



of the estimator. In our implementation of our estimator, we have made a further simplification by using the fact that (5) can be explicitly solved in case of the Frank copula. The resulting computing time for one estimation with 121 time points and sample size 50, 500 and 1000 is about 48 seconds, 7.4 minutes and 14.4 minutes respectively. These numbers are obtained with a Quad-Core Xeon 2.66 GHz with Matlab for Linux. This gives further evidence for the complexity of the underlying numerical problem.

Table 4: Mean squared error of the estimated treatment effect

Sample Size		50			500			1000		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Risk 1	$t = 0.07$	0.0015	0.0051	0.0051	0.0001	0.0005	0.0005	0.0006	0.0002	0.0002
	$t = 0.24$	0.0106	0.0096	0.0097	0.0023	0.0010	0.0010	0.0030	0.0005	0.0005
	$t = 1.20$	0.0280	0.0252	0.0260	0.0121	0.0026	0.0027	0.0126	0.0013	0.0015
Risk 2	$t = 0.07$	0.0013	0.0035	0.0035	-0.0011	0.0003	0.0003	-0.0011	0.0002	0.0002
	$t = 0.24$	-0.0034	0.0106	0.0106	-0.0023	0.0010	0.0010	-0.0047	0.0006	0.0006
	$t = 1.20$	-0.0165	0.0169	0.0171	-0.0152	0.0017	0.0019	-0.0117	0.0008	0.0009
Risk 3	$t = 0.07$	0.0009	0.0023	0.0023	-0.0007	0.0002	0.0002	-0.0007	0.0001	0.0001
	$t = 0.24$	-0.0013	0.0093	0.0093	-0.0009	0.0010	0.0010	-0.0015	0.0004	0.0004
	$t = 1.20$	0.0032	0.0422	0.0422	0.0077	0.0034	0.0035	0.0034	0.0018	0.0018

This section has demonstrated the applicability of the risk pooling method with simulated data. The results of our simulation study confirm the nice statistical properties of the copula

based estimator. As a next step we put the estimator to real world data.

4 Application

The last section has shown that the copula based estimator produces good results for a model with more than two risks. It has also shown that a misspecified dependence structure between competing risks can produce rather biased result patterns. In this section we apply the copula based estimator to real world unemployment duration data from Germany to check the robustness of empirical results with respect to the assumed dependence structure. We analyse the effects of a policy reform conducted in the year 1997 which decreased the entitlement length for unemployment benefits (UB) for older unemployed. Younger unemployed were not directly affected by the reform and our statistical evaluation concept assumes the absence of indirect effects as they are supposedly small. The reform was therefore a natural experiment and we apply a difference in differences (DID) approach for the estimation of the treatment effect on survival probabilities in unemployment. Our control group consists of the younger unemployed (aged 36-41), while the older unemployed (aged 42-44) build the treatment group. As a result of the policy change, the UB entitlement lengths for the latter group decreased from up to 22 months to 12 months. We take the period 1995-1996 as pre-reform and 1999-2000 as post reform period. For the estimations we use a sample of the IAB employment sample 2001 of the Institute for Employment Research (IAB), Nuremberg. The data is a 2% random sample of the German workforce subject to social security contributions in the period 1975-2001. It contains daily information about periods of dependent employment and claim periods for unemployment benefits. Moreover, it contains basic information about the individual, job, employer and regional characteristics. For more details about this data see Hamann et. al. (2004). In a recent paper Arntz et al. (2007) already analyse the effect of the above mentioned reform using the same sample of observations. They estimate the difference in differences changes in nonparametric risk specific CIC's. Their model allows for transitions to three risks: local employment, distant employment and unknown exit. Distant employment is usually linked with an inter-regional migration decision of the unemployed. With this model it is therefore possible to analyse the theoretical question whether unemployment benefits simply decrease the job finding hazard or whether they provide resources for the unemployed to migrate to distant areas and thus increases the hazard of distant job finding. In their paper it is argued that the theoretical predictions for the effect of UB entitlement lengths on distant employment timings are unclear and their work aims to analyse this empirically. Given the non-identifiability of the competing risks model, they face the problem that they can only consistently estimate the joint survival function and the risk specific CICs as they are not willing to impose further identifying

assumptions. For this reason they cannot estimate the DID changes in the marginal distributions. As an additional complication, the true unemployment duration is not observed in the IAB data. For this reason they use an upper bound and a lower bound of the true unemployment duration which can be determined by the data. They bound the estimated policy effect on the risk specific CICs by exploiting that the nonparametric estimates posses a monotonic relation between the bounds of the duration. We are ignoring the partial identification of unemployment duration in our following analysis by using one bound of the unemployment duration only. This helps us in focusing on the applicability and benefits of the copula model, although an incorporation of the partial identification problem would be generally possible. In particular, we use their sample of lower bounds of the true unemployment duration. Moreover, we restrict our analysis to the group of higher skilled males because we are expecting that the decrease in UB entitlement lengths has a stronger financial effect for this group. Our sample consists of 2,095 observations.

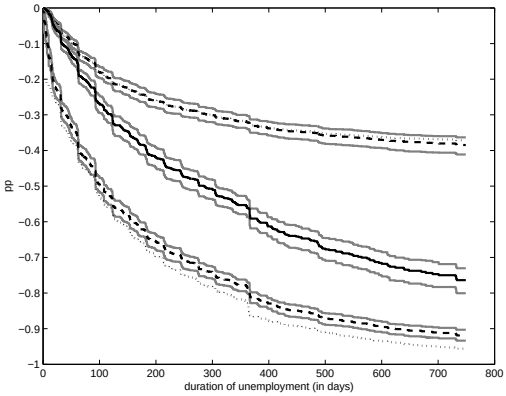
As we are not aware of any economic theory which suggests a dependence structure between competing risk under plausible assumptions, we are facing the problem that it is unknown to us. For this reason we are not able to identify the *true* treatment effect of the reform for the given definition of unemployment. Although it is impossible to break up the fundamental identification problem, we reason that by using the copula based estimator, we obtain valuable information for assessing the sensitivity of empirical result patterns with respect to the assumed dependence structure. If the empirical results are robust one can conclude that a simple Kaplan-Meier estimator would make a good job. If the sign of the estimated treatment effect is not robust, one can probably not draw any causal conclusions for the changes in the marginal distributions in absence of knowledge about the dependence structure. When putting the estimator to data, we observe similar to Zheng and Klein (1995) that the choice of the copula is less relevant for the results than the choice of its parameter. For this reason we report results for the one copula with different parameter values only. In particular, we choose the Frank copula with parameter τ, the so called Kendall's τ, which measures the dependence degree between the competing risks.

Figure 2 presents the estimated marginal survival functions for the risks local employment and distant employment. Since the distribution for unknown exit is not meaningful, we do not report results for it. We compare the copula based estimator with three different parameters to the KM estimator and the Peterson bounds. As the KM estimator is very close to the copula based estimator with assumed independence ($\tau = 0$), we do not report the latter. In both cases, the estimated marginal survivors strongly vary with the assumed dependence structure. They differ by up to 40 percentage points after two years of unemployment. Knowledge of the true dependence structure seems to be important for the interpretation of the resulting estimates. However the shape of the marginal survivors is similar for different assumed dependence structures. As a next

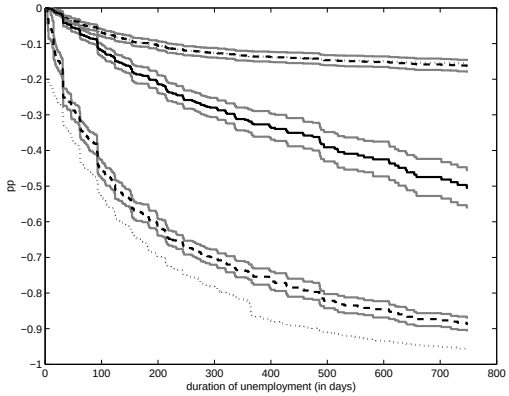
Figure 2: Estimated marginal survival functions with different dependence structures: KM Estimator(solid line); τ = -0.8 (upper dashed line); $\tau = 0.8$ (lower dashed line); 95% and 5% bootstrap quantiles (grey dashed lines); Peterson bounds (dotted line)

Aged < 42, before 1997

(a) Local employment



(b) Distant employment



step we estimate the difference in differences treatment effect on these marginal survival functions.

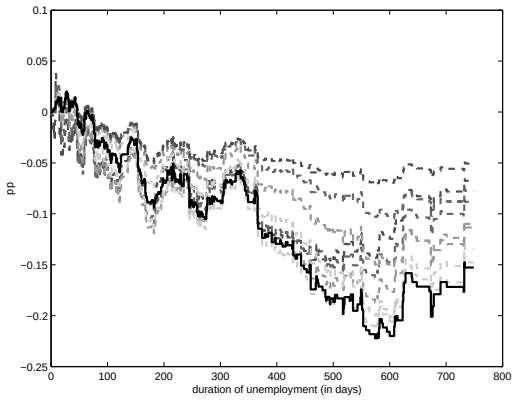
First we assume that the dependence structure is the same for all groups, i.e. it does not change in response to the reform. Figure 3 presents the estimated treatment effect for different copula parameter values. It shows that the estimated treatment effects have a similar shape. The estimates are more similar for short durations, while the shape of the estimated treatment effect is more sensitive to the specification of the dependence structure for long durations. This is in particular the case for distant employment. Since the reform is likely to affect longer durations (between days 365 and 600), the specification of the dependence structure seems to be a relevant issue for the evaluation of the reform effect. However, the sign of the effect is in most cases independent of the dependence structure. Also note that the KM estimator does not necessarily lie between the other estimates.

As a next step we explore the results in case the dependence structure varies in response to the reform. Figure 4 presents the results for a constant dependence structure in pre-reform period and different dependencies after the reform. As in the previous case, the KM estimator does not lie between different copula based estimates. The sign of the estimated treatment effect now depends more often on the assumed dependence structure, but it is quite robust.

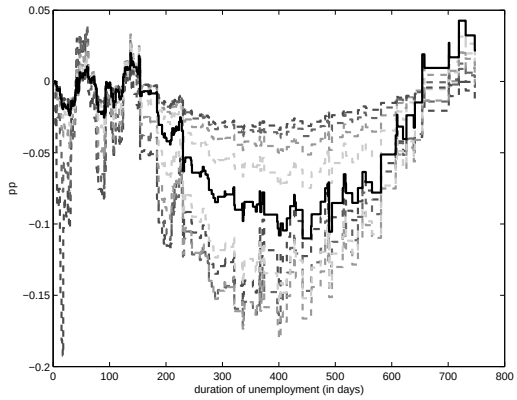
The results with constant dependence structure are reported on the diagonal of Table 5. The lower triangle of this table corresponds to the case when the reform decreases τ, while the entries above the diagonal refer to the case when τ increases in response to the reform. The results suggest that the KM estimator differs in some cases considerably from the results obtained with

Figure 3: Estimated treatment effect with different dependence structures. KM Estimator (dark line); estimate with $\tau \in [-0.8, 0.8]$ (grey dashed line in different darkness)

(a) Local employment



(b) Distant employment

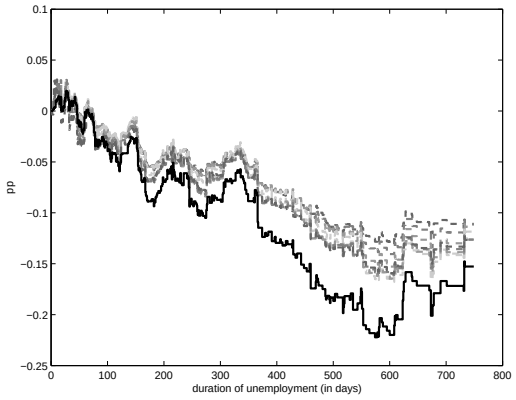


other dependence structures. There is no apparent systematic pattern when the KM estimator under- or over-estimates the reform effect.

Our application shows that the KM estimator is able to produce a quite robust estimate for the sign and the general pattern of the treatment effect, while the magnitude is often rather misleading. The empirical results suggest that the marginal distributions of both risks shifted to the left in response to the reform. This is also confirmed when we repeat the estimations for the upper bound of the true unemployment duration. Our empirical exercise has shown that the copula based estimator is applicable to applied economic problems. It is very powerful if one has information about the dependence structure in an application. Even in the case when such information is not available, it is a helpful tool to check the sensitivity of results with respect to the assumed dependence structure.

Figure 4: Estimated treatment effect with changing dependence structure. Before the reform, τ_0 = -0.4. After the reform, $\tau_1 \in [-0.8, 0.8]$. KM Estimator ($\tau_0 = \tau_1 = 0$, dark line); estimates with different post reform τ_1 (grey dashed line in different darkness)

(a) Local employment



(b) Distant employment

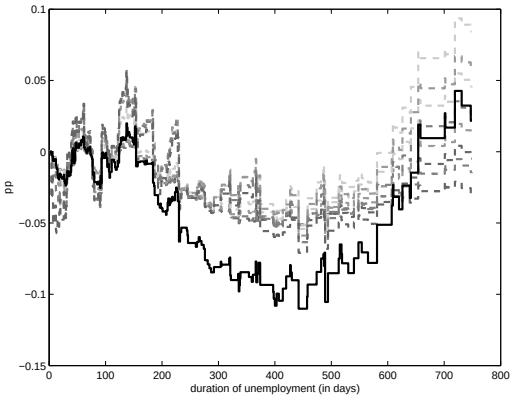


Table 5: Sensitivity analysis of estimated treatment effect with different dependence structure before (rows) and after (columns) the reform. Dependence is measured by Kendall's tau.

(a) Local

t=180, KM=-0.0868

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0444	-0.0448	-0.0490	-0.0561	-0.0576
-0.4	-0.0533	-0.0536	-0.0579	-0.0649	-0.0665
0.0	-0.0889	-0.0892	-0.0935	-0.1005	-0.1021
0.4	-0.0994	-0.0998	-0.1040	-0.1111	-0.1126
0.8	-0.0747	-0.0750	-0.0793	-0.0863	-0.0879

t=360, KM=-0.0856

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0368	-0.0354	-0.0404	-0.0453	-0.0458
-0.4	-0.0542	-0.0529	-0.0578	-0.0627	-0.0632
0.0	-0.0861	-0.0847	-0.0896	-0.0945	-0.0950
0.4	-0.0779	-0.0766	-0.0815	-0.0864	-0.0869
0.8	-0.0675	-0.0661	-0.0711	-0.0760	-0.0765

t=550, KM=-0.1743

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0567	-0.0620	-0.0734	-0.0764	-0.0784
-0.4	-0.1077	-0.1130	-0.1244	-0.1274	-0.1294
0.0	-0.1599	-0.1652	-0.1765	-0.1795	-0.1816
0.4	-0.1323	-0.1376	-0.1490	-0.1520	-0.1540
0.8	-0.0970	-0.1023	-0.1137	-0.1167	-0.1187

(b) Distant

t=180, KM=-0.0084

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0031	-0.0008	0.0115	0.0252	0.0177
-0.4	-0.0061	-0.0038	0.0084	0.0221	0.0146
0.0	-0.0250	-0.0227	-0.0105	0.0032	-0.0043
0.4	-0.0560	-0.0537	-0.0414	-0.0277	-0.0352
0.8	-0.0714	-0.0692	-0.0569	-0.0432	-0.0507

t=360, KM=-0.0913

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0319	-0.0265	-0.0088	-0.0027	-0.0140
-0.4	-0.0499	-0.0445	-0.0268	-0.0207	-0.0320
0.0	-0.1282	-0.1229	-0.1052	-0.0991	-0.1104
0.4	-0.1941	-0.1888	-0.1710	-0.1649	-0.1762
0.8	-0.1661	-0.1608	-0.1430	-0.1369	-0.1482

t=550, KM=-0.0635

	-0.8	-0.4	0.0	0.4	0.8
-0.8	-0.0181	-0.0111	0.0036	-0.0049	-0.0262
-0.4	-0.0366	-0.0296	-0.0149	-0.0234	-0.0446
0.0	-0.0960	-0.0889	-0.0743	-0.0828	-0.1040
0.4	-0.1055	-0.0985	-0.0838	-0.0923	-0.1135
0.8	-0.0621	-0.0551	-0.0404	-0.0490	-0.0702

5 Conclusion

We adapt the Copula Graphic Estimator of Zheng and Klein to the case of more than two dependent competing risks if the copula function belongs to the Archimedean family. Our implementation works with common data structures as it is for example compatible with non-distinct observations. We obtain evidence that our estimator is an interesting alternative to the Kaplan-Meier estimator as our simulations and our application demonstrate its applicability. In contrast to the Kaplan-Meier estimator, the copula based estimator does not require independence of competing risks but it requires an assumption about the basic dependence structure between competing risks. Unfortunately, the latter cannot be tested and the choice of the dependence structure between competing risks is therefore a non-testable identifying assumption even if it is derived from economic theory. The copula approach is therefore not able to break up the non identification of the competing risks model since resulting marginal distributions can attain basically any point within the nonparametric Peterson bounds. However, this is not a particular weakness of this approach as the identification of other popular duration models such as the proportional hazard model or mixed proportional hazard model is simply achieved due to more restrictive model assumptions. One can for example show that a special case of the copula model is a mixed proportional hazard model.

References

[1] Abbring, J.H. and G.J. van den Berg (2003) The identifiability of the mixed proportional hazards competing risks model, *Journal of the Royal Statistical Society B*, 65, 701–710.

[2] Amemiya, T. (1985) *Advanced Econometrics*, Oxford: Basil Blackwell.

[3] Arntz, M., Lo, S. and Wilke, R.A. (2007) Bounds analysis of competing risks: a nonparametric evaluation of the effect of unemployment benefits on migration in Germany, *ZEW Discussion Paper No. 07-049*. ZEW, Mannheim.

[4] Braekers, R. and Veraverbeke, N. (2006) A Copula-Graphic Estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, 33, 429–447.

[5] Carrière, J.F. (1995) Removing Cancer when it is Correlated with other Causes of Death. *Biometric Journal*, 3, 339–350.

[6] Clayton, D.G. (1978) A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence, *Biometrika*, 65, 141-151.

[7] Cox, D.R. (1962) *Renewal Theory*, London.

[8] Genest, C., Ghoudi, K. and Rivest, L.P. (1995) A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions, *Biometrika*, 82, 543–552.

[9] Hamann, S., G. Krug, M. Köhler, W. Ludwig-Mayerhofer, and A. Hacket (2004) Die IAB-Regionalstichprobe 1975-2001: IABS-R01, *ZA-Information* 55, 36–42.

[10] Heckman, J.J. and B.E. Honoré (1989) The identifiability of the competing risks model, *Biometrika*, 76, 325–330.

[11] Horowitz, J. L. and Manski, C. F. (2000) Nonparametric analysis of randomized experiments with missing covariate and outcome data (with discussion), *J. Amer. Statist. Assoc.*, 95, 77–88.

[12] Joe, H. (1994) Multivariate Extreme-Value Distributions with Applications to Environmental Data. *Canadian Journal of Statistics*, 22, 47–64.

[13] Kalbfleisch, J.D. and R.L. Prentice (1980) The Statistical Analysis of Failure Time Data, *Wiley Series in Probability and Statistics*.

[14] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, 53, 457–481.

[15] Mikosch, T. (2006) Copulas: Tales and facts, *Extremes*, 9, 3–20.

[16] Nelsen, R.B. (2006) *An Introduction to Copulas*, 2nd Edition, Springer, New York.

[17] Oakes, D. (1989) Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84, 487–493.

[18] Peterson, A.V. (1976) Bounds for a Joint Distribution With Fixed Sub-Distribution Functions: Application to Competing Risks, *Proceedings of the National Academy of Science*, 73, 11–13.

[19] Rivest, L.P. and Wells, M.T. (2001) A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring, *Journal of Mulitvariate Analysis*, 79, 138-155.

[20] Schweizer, B. and Sklar, A. (1983) *Probabilistic Metric Spaces*, New York: North-Holland.

[21] Tsiatis, A. (1975) A Nonidentifiability Aspect of the Problem of Competing Risks, *Proceedings of the National Academy of Sciences*, 72, 20–22.

[22] Van den Berg, G. (2001) Duration Models: Specification, Identification and Multiple Durations. In *Handbook of Econometrics*, Vol.5, North-Holland.

[23] Zheng, M. and Klein, J.P. (1995) Estimates of marginal survival for dependent competing risks based on assumed copula. *Biometrika*, 82, 127-138.

[24] Zimmer, D.M. and Trivedi, P.K. (2006) Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand. *Journal of Business and Economic Statistics*, 24, 63–76.

Imprint

FDZ Methodenreport
No. 02/2009

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Service
at the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Herrlinger

Technical production
Dagmar Herrlinger

Download
http://doku.iab.de/fdz/reporte/2009/MR_02-09.pdf

Internet
http://fdz.iab.de/

Corresponding author
Ralf A. Wilke
University of Nottingham
E-Mail: ralf.wilke@nottingham.ac.uk