Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size

PRENEN, Leen; BRAEKERS, Roel & DUCHATEAU, Luc (2017) Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. In: Journal of the Royal Statistical Society. Series B: Statistical methodology, 79 (2), p. 483-505.

DOI: 10.1111/rssb.12174
Handle: http://hdl.handle.net/1942/21320

Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size

Leen Prenen[1], Roel Braekers *[1] and Luc Duchateau[2]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium

[2]Department of Physiology and Biometrics, Universiteit Gent, Belgium

Abstract

For the analysis of clustered survival data, two different types of models that take the association into account, are commonly used: frailty models and copula models. Frailty models assume that conditional on a frailty term for each cluster, the hazard functions of individuals within that cluster are independent. These unknown frailty terms with their imposed distribution are used to express the association between the different individuals in a cluster. Copula models on the other hand assume that the joint survival function of the individuals within a cluster is given by a copula function, evaluated in the marginal survival function of each individual. It is the copula function which describes the association between the lifetimes within a cluster. A major disadvantage of the present copula models over the frailty models is that the size of the different clusters must be small and equal in order to set up manageable estimation procedures for the different model parameters. We describe in this manuscript a copula model for clustered survival data where the clusters are allowed to be moderate to large and varying in size by considering the class of Archimedean copulas with completely monotone generator. We develop both one- and two-stage estimators for the different copula parameters. Furthermore we show the consistency and asymptotic normality of these estimators. Finally, we perform a simulation study to investigate the finite sample properties of the estimators. We illustrate the method on a data set containing the time to first insemination in cows, with cows clustered in herds.

Keywords: Archimedean copula, multivariate survival data, varying cluster size

*Address: Roel Braekers, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium. E-mail: roel.braekers@uhasselt.be

1 Introduction

Multivariate survival data consist of multiple lifetimes which are linked to each other in some sense. In clustered survival data, subjects in the same cluster are assumed to share some characteristic or environment, and are therefore expected to be more similar with respect to the hazard of the event. For example, in a multi-center clinical trial, patients of one center form a separate cluster. To analyze this type of multivariate survival data, two different techniques that take the association between the individuals into account, are commonly used, namely frailty models and copula models.

In frailty models, the interest lies on the hazard function of an individual, conditionally on an unknown frailty term for the cluster containing this individual. In these models, we follow a conditional viewpoint and investigate the influence of different covariates on the hazard function of an individual, given the cluster. The frailty term for each cluster expresses that we assume that different individuals in the same cluster behave in a similar but unknown manner. We consider this frailty term as a realization of a random variable with a given frailty distribution and allow it to vary over the different clusters. This approach is explained in detail in Duchateau and Janssen (2008) and Wienke (2011).

To estimate the different parameters in frailty models, we make use of the conditional viewpoint of these models. Hereby we assume that different individuals within the same cluster are treated as independent of each other, conditionally on this common frailty term. In the construction of the likelihood function of a frailty model, this assumption is utilized by first looking at the conditional contribution of an individual within a cluster to the likelihood function and afterwards integrating over the frailty distribution. In this way, the frailty model approach has the advantage that it allows that the number of individuals within a cluster may vary over the different clusters. However, a major disadvantage of the frailty model is that the marginal survival functions in the frailty model contain the association parameter of the frailty distribution (Goethals et al. (2008)). This has led to the correct observation by, e.g., (Hougaard, 1986, p. 676) that the association parameter in a frailty model can be obtained from the marginal survival functions alone. Additionally, overdispersion in the data, as compared to the proposed density function, is required in a frailty model in order to pick up association.

Copula models, on the other hand, are specified in terms of the marginal distribution of an individual. The association between different individuals within a cluster is modelled by introducing a copula function that links the marginal survival functions together to obtain the joint survival function.

To estimate the different parameters in copula models, often two stages are used. In the first stage, the parameters of the marginal survival functions are estimated, and then inserted in the copula function. In the second stage, the parameter(s) of the copula function are estimated.

Thus, both in the model specification and parameter estimation, the parameter(s) describing the association is kept separate from the other parameters. Most reported copula models, however, only use clusters in which the cluster size is small and constant over the different clusters as it is then straightforward to define and estimate the marginal survival functions. For example, Shih and Louis (1995) introduced a copula model for multivariate survival data and provided estimation methods for the unknown parameters in a bivariate setting. Glidden (2000) and Andersen (2005) extended the approach of Shih and Louis (1995) to include covariates into the marginal survival function, but also here the clusters only had size two. Massonnet et al. (2009) extended these models further for clusters of size 4 to model the time until infection in the four different quarters of a cow udder. Although Glidden (2000) gives theoretical results for the Clayton copula in a balanced design with a fixed cluster size N and Othus and Li (2010) do the same in an unbalanced design for the Gaussian copula model, to our knowledge, copula models in general have not been used for clustered multivariate survival data with a cluster size of more than 4 or for a cluster size which differs over the clusters. The choice of a small and constant cluster size is a direct consequence of the difficulty to write down the likelihood function for the observed clustered survival data. For example, if the cluster size is equal to two, there are 4 different contributions to the likelihood for the observed outcomes within the cluster, depending on whether none, the first, the second or both individuals in this cluster are censored. This leads to a likelihood function consisting of 4 different terms where every term is found by taking derivatives of the joint survival function over the uncensored components in an observed couple. If the cluster size is three, the number of possible combinations increases to 8, while a cluster size of 4 leads to 16 different combinations. In a general setting with a cluster size equal to n, we have 2^n possible combinations. Since a likelihood function also contains 2^n different possible terms and each term is found by taking derivatives of the joint survival function over the uncensored components in a combination, it is a huge task to get an expression for the likelihood function when a general n-dimensional copula function is considered for the association between the different individuals within a cluster. In practice it is impossible to calculate a closed form for all the derivatives of a copula function if the order n is large.

For the class of Archimedean copula functions, we will solve this numerical problem in this manuscript and show that the construction of the likelihood function for this class of copula functions simplifies considerably such that we can allow the cluster size to be moderate to large and varying over the different clusters. The key to this solution is that the joint survival function of an Archimedean copula function can be rewritten as a mixture distribution of independent contributions in a similar way as in the frailty model approach. Although some of the expressions of the Archimedean copula function resemble that of the frailty model, the two models differ in an essential way due to their different inferential viewpoint, i.e., marginal versus conditional.

The article is organized as follows. In Section 2 we introduce a new formulation of the Archimedean copula model by rewriting the likelihood contributions in terms of Laplace transforms. In Section 3 we present the theoretical results concerning estimators arising from this model, starting

from parametric and semiparametric approaches. Section 4 gives an overview of a large class of distributions for which the likelihood contributions are easy to generate. In Sections 5 and 6, we report simulation results along with results for a data example. The data set and our code can be found at our website (http://www.vetstat.ugent.be/research/ArchimedeanCopula/). Proofs of asymptotic results are given in the Appendix.

2 Description of the model

We develop a copula model for clustered survival data in which the size of each cluster may be different. Let K be the number of clusters ($i = 1, \ldots, K$). In each cluster, we denote the lifetime for the different individuals by a positive random variable T_{ij}, $j = 1, \ldots, n_i$ where n_i is the number of individuals in cluster i. For each individual, we assume that there is an independent random censoring variable C_{ij} such that under a right censoring scheme, the observed quantities are given by

$$\begin{array}{l} X_{ij} = \min(T_{ij}, C_{ij}) \\ \delta_{ij} = I(T_{ij} \leq C_{ij}) \end{array} \quad , i = 1, \ldots, K, \quad j = 1, \ldots, n_i.$$

The risk of failure may also depend on a set of covariates $\boldsymbol{Z}_{ij}$, which are possibly time-varying. We assume that the joint survival function for the lifetime of the different individuals within cluster i is given by

$$\begin{aligned} S(t_{i1}, \ldots, t_{in_i} | \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i}) &= P(T_{i1} > t_{i1}, \ldots, T_{in_i} > t_{in_i} | \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i}) \\ &= \varphi_\theta \left[\varphi_\theta^{-1} \left(S(t_{i1} | \boldsymbol{Z}_{i1}) \right) + \cdots + \varphi_\theta^{-1} \left(S(t_{in_i} | \boldsymbol{Z}_{in_i}) \right) \right] \end{aligned}$$

where $S(t_{ij}|\boldsymbol{Z}_{ij}) = P(T_{ij} > t_{ij}|\boldsymbol{Z}_{ij})$ is a common marginal survival model for the lifetime T_{ij}, given $\boldsymbol{Z}_{ij}$. The generator $\varphi_\theta : [0, \infty[\rightarrow [0, 1]$ of a parametric Archimedean copula family is a continuous strictly decreasing function with $\varphi_\theta(0) = 1$ and $\varphi_\theta(\infty) = 0$. We denote by φ_θ^{-1} the inverse function of φ_θ. Since we want the Archimedean copula function to be correctly defined for any cluster size, we assume that this generator is completely monotonic. This means that all the derivatives exist and have alternating signs: $(-1)^m \frac{d^m}{dt^m} \varphi_\theta(t) \geq 0$, for all $t > 0$ and $m = 0, 1, 2, \ldots$ (see Nelsen (2006)). The generator φ_θ is a Laplace transformation of a positive distribution function $G_\theta(x)$ with $\bar{G}_\theta(0) = 1$ (Joe, 1997),

$$\varphi_\theta(t) = \int_0^{+\infty} e^{-tx} dG_\theta(x), \quad t \geq 0.$$

Hence we can rewrite the joint survival function for cluster i as

$$\begin{aligned} S(t_{i1}, \ldots, t_{in_i} | \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i}) &= \int_0^{+\infty} e^{-x \sum_{j=1}^{n_i} \varphi_\theta^{-1}(S(t_{ij}|\boldsymbol{Z}_{ij}))} dG_\theta(x) \qquad (1) \\ &= \int_0^{+\infty} \prod_{j=1}^{n_i} e^{-x \varphi_\theta^{-1}(S(t_{ij}|\boldsymbol{Z}_{ij}))} dG_\theta(x). \end{aligned}$$

In this way, the Archimedean copula function can be seen as a mixture distribution, consisting of independent and identically distributed components which depend on a common factor that has G_θ as distribution. We use this structure to derive the likelihood function. The contribution of cluster i, with cluster size n_i, to the likelihood function corresponds to the derivative of the n_i-dimensional joint survival function over all uncensored individuals in this cluster. Since the joint survival function does not change when the individuals within the cluster are permuted, we note that only the number of uncensored individuals determines the derivative. Hence, the contribution of cluster i to the likelihood function is given by

$$L_i = (-1)^{d_i} \frac{\partial^{d_i}}{\partial\{\delta_{ij}=1\}} S(x_{i1}, \ldots, x_{in_i} | \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i})$$

where $\partial\{\delta_{ij}=1\}$ is the set of uncensored individuals in cluster i and $d_i = \sum_{j=1}^{n_i} \delta_{ij}$, the size of this set.

Using representation (1) of the joint survival function, this derivative is given by

$$L_i = \int_0^{+\infty} e^{-x \sum_{j=1}^{n_i} \varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij}))} \prod_{j=1}^{n_i} \left[\frac{-x f(x_{ij}|\boldsymbol{Z}_{ij})}{\varphi'_\theta(\varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij})))} \right]^{\delta_{ij}} dG_\theta(x)$$

where $f = -dS/dt$ is the conditional density of the lifetime X_{ij}.

Combining the contributions over the different clusters, we get the following likelihood function

$$\begin{aligned} L &= \prod_{i=1}^{K} \int_0^{+\infty} e^{-x \sum_{j=1}^{n_i} \varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij}))} \prod_{j=1}^{n_i} \left[\frac{-x f(x_{ij}|\boldsymbol{Z_{ij}})}{\varphi'_\theta(\varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z_{ij}})))} \right]^{\delta_{ij}} dG_\theta(x) \\ &= \prod_{i=1}^{K} \int_0^{+\infty} \prod_{j=1}^{n_i} e^{-x\varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij}))} \left[\frac{-x f(x_{ij}|\boldsymbol{Z}_{ij})}{\varphi'_\theta(\varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z_{ij}})))} \right]^{\delta_{ij}} dG_\theta(x). \end{aligned} \tag{2}$$

In general it is difficult to evaluate expression (2) except for very specific choices of the distribution G_θ. Since the generator φ_θ is the Laplace transform of G_θ, there is an alternative expression for this likelihood function which is found by using derivatives of this generator, i.e. $\varphi_\theta^{(m)}(t) = \int_0^{+\infty} (-x)^m e^{-tx} dG_\theta(x)$. Hence the likelihood function can be rewritten as

$$L = \prod_{i=1}^{K} \left(\prod_{j=1}^{n_i} \left[\frac{f(x_{ij}|\boldsymbol{Z}_{ij})}{\varphi'_\theta(\varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij})))} \right]^{\delta_{ij}} \right) \varphi_\theta^{(d_i)} \left(\sum_{j=1}^{n_i} \varphi_\theta^{-1}(S(x_{ij}|\boldsymbol{Z}_{ij})) \right). \tag{3}$$

Remark: In the frailty model framework (Duchateau and Janssen, 2008, p.119), we note that we find a similar expression for the joint survival function in frailty models, with $G_\theta(x)$ as the frailty distribution of the unknown frailty term in the cluster. Starting from the conditional viewpoint in frailty models, we find a similar expression for the joint survival function as follows.

The joint conditional survival function for a cluster i is given by $S(t_{i1},\ldots,t_{in_i}|Z_{i1},\ldots,Z_{in_i},U_i)$ with U_i the frailty term with distribution $G_\theta(u)$ and generator $\varphi_\theta(\cdot)$. Denote the conditional cumulative hazard function for subject j from cluster i by $H(t_{ij}|Z_{ij},U_i) = H_c(t_{ij}|Z_{ij})U_i$. The marginal joint survival function is obtained by integrating out the frailty term:

$$\begin{aligned} S_f(t_{i1},\ldots,t_{in_i}|Z_{i1},\ldots,Z_{in_i}) &= \int_0^\infty S(t_{i1},\ldots,t_{in_i}|Z_{i1},\ldots,Z_{in_i},u_i)dG_\theta(u_i) \\ &= \int_0^\infty S(t_{i1}|Z_{i1},u_i)\ldots S(t_{in_i}|Z_{in_i},u_i)dG_\theta(u_i) \\ &= \int_0^\infty \exp(-u_i\sum_{j=1}^{n_i} H_c(t_{ij}|Z_{ij}))dG_\theta(u_i) \\ &= \int_0^\infty \exp(-u_i\sum_{j=1}^{n_i} \varphi_\theta^{-1}(S_f(t_{ij}|Z_{ij})))dG_\theta(u_i) \end{aligned} \quad (4)$$

due to the conditional independence assumption. The two joint survival functions (1) and (4) are indeed similar, but note that $S(t_{ij}|Z_{ij}) \neq S_f(t_{ij}|Z_{ij})$. More specifically $S_f(t_{ij}|Z_{ij}) = \varphi_\theta(H_c(t_{ij}|Z_{ij}))$ and therefore, the marginal survival function in (4) contains the association parameter. This an important distinction between the frailty model and the copula model.

3 The estimation procedures

In this section, we investigate a one- and two-stage parametric estimation method and a two-stage semi-parametric estimation method to estimate the different parameters in this model. Shih and Louis (1995) demonstrated how this can be done for a bivariate survival data set and derived asymptotic properties of the estimators. Joe (1997, 2005) discussed a general framework for studying asymptotic efficiency. We extend their results to clustered survival data with clusters of varying and possibly large size.

For equal-sized clusters with cluster size n having the same covariate structure, baseline survival functions can be estimated for each j^{th} univariate margin, $j = 1,\ldots,n$, where the j^{th} subject always has the same covariate information. Since in our application clusters have varying size, we cannot order the components within a cluster and estimate the baseline survival of all j^{th} components. We assume that all subjects have the same baseline survival, whatever the cluster, and introduce subject specific covariate information.

3.1 One-stage parametric estimation

Let $\boldsymbol{\beta}$ be the parameter vector for the margins, containing distribution-specific parameters for the baseline survival and covariate effects. We use the likelihood function $L(\boldsymbol{\beta}, \theta)$ as derived in (2) and (3). Write $\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \theta) = \frac{\partial \log L(\boldsymbol{\beta},\theta)}{\partial \boldsymbol{\beta}}, U_\theta(\boldsymbol{\beta}, \theta) = \frac{\partial \log L(\boldsymbol{\beta},\theta)}{\partial \theta}$. Solving

$$\begin{cases} \mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \theta) = 0 \\ U_\theta(\boldsymbol{\beta}, \theta) = 0 \end{cases}$$

simultaneously, we find the maximum likelihood estimate $(\hat{\boldsymbol{\beta}}, \hat{\theta})$. From maximum likelihood theory (Cox and Hinkley, 1974), we know that under regularity conditions, $\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{\theta} - \theta)$ converges to a multivariate normal distribution with mean vector zero and variance-covariance matrix $\mathbf{I}^{-1}$, where $\mathbf{I}$ is partitioned into blocks:

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{I}_{\boldsymbol{\beta}\theta} \\ \mathbf{I}_{\theta\boldsymbol{\beta}} & I_{\theta\theta} \end{pmatrix}.$$

Here, $K\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is the variance-covariance matrix of $\mathbf{U}_{\boldsymbol{\beta}}$, $K\mathbf{I}_{\boldsymbol{\beta}\theta}$ is the covariance vector between $\mathbf{U}_{\boldsymbol{\beta}}$ and U_θ and $KI_{\theta\theta}$ is the scalar variance of U_θ, so

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{I_{\theta\theta}} + \frac{\mathbf{I}_{\theta\boldsymbol{\beta}}(\mathbf{I}^{-1})_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{I}_{\boldsymbol{\beta}\theta}}{I_{\theta\theta}^2}. \tag{5}$$

In practical applications, standard errors of parameter estimates can be retrieved from the diagonal elements of the inverse of the Hessian matrix $\mathbf{I}$.

3.2 Two-stage parametric estimation

Two-stage parametric estimation, also referred to as the method of inference functions for margins (Xu, 1996), has been used mainly for multivariate models whenever a multi-parameter numerical optimization for maximum likelihood estimation is too time-consuming or infeasible. In the first stage, $\boldsymbol{\beta}$ is estimated by $\overline{\boldsymbol{\beta}}$ by considering all subjects as independent, identically distributed random variables, i.e. solving

$$\mathbf{U}^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \delta_{ij}\frac{\partial \log f(x_{ij}|\boldsymbol{Z}_{ij})}{\partial \boldsymbol{\beta}} + (1-\delta_{ij})\frac{\partial \log S(x_{ij}|\boldsymbol{Z}_{ij})}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Under regularity conditions, $\sqrt{K}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a multivariate normal distribution with mean vector zero and variance-covariance matrix $(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}$, where $\mathbf{V}$ is the variance-covariance matrix of the score functions $\mathbf{U}^*_{\boldsymbol{\beta}}$ and $\mathbf{I}^*$ is the Fisher information of $\mathbf{U}^*_{\boldsymbol{\beta}}$. The use of the robust sandwich estimator is required since $(\mathbf{I}^*)^{-1}$ is not a consistent estimator of the asymptotic variance-covariance matrix due to the correlation between survival times. In the

second stage, the association parameter θ is estimated by plugging in the estimates for the margins into the likelihood expression (3), which is then maximized for the association parameter θ. The two-stage estimator for θ is the solution to

$$U_\theta(\overline{\boldsymbol{\beta}}, \theta) = \frac{\partial \log L}{\partial \theta}(\overline{\boldsymbol{\beta}}, \theta) = 0.$$

Theorem 1. *Let $\overline{\theta}$ denote the solution to $U_\theta(\overline{\boldsymbol{\beta}}, \theta) = 0$ and let θ_0 be the true value of the association parameter. Under regularity conditions, $\sqrt{K}(\overline{\theta} - \theta_0)$ converges to a normal distribution with mean zero and variance*

$$Var(\overline{\theta}) = \frac{1}{I_{\theta\theta}} + \frac{\mathbf{I}_{\theta\beta}(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}\mathbf{I}_{\beta\theta}}{I_{\theta\theta}^2}. \tag{6}$$

The proof of Theorem 1 is provided in the Appendix. To estimate this quantity, we make use of $(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}$, the robust variance obtained in the first step; $I_{\theta\theta}^{-1}$ and $I_{\beta\theta}$ are obtained from the Hessian matrix of the one-stage procedure, which can be estimated numerically by performing one iteration of the one-stage optimization in which we evaluate the Hessian matrix in the two-stage parameter results.

3.3 Two-stage semiparametric estimation

In the two-stage semiparametric estimation procedure, the marginal survival functions are estimated using the Cox proportional hazards model (Cox, 1972). Formulas for the standard error of the estimated covariate effect $\check{\boldsymbol{\beta}}$ and the estimated cumulative hazard $\check{\Lambda}$ that account for clustering can be found using a sandwich formula (Spiekerman and Lin, 1998).

In the second stage, $\max_\theta L(\theta; \check{\boldsymbol{\beta}}, \check{\Lambda})$ is solved for $\check{\theta}$.

Theorem 2. *Under regularity conditions C.1-C.7 in the Appendix, $(\check{\theta}; \check{\boldsymbol{\beta}}, \check{\Lambda})$ is a consistent estimator for $(\theta_0; \boldsymbol{\beta}_0, \Lambda_0)$.*

The results for $\check{\boldsymbol{\beta}}$ and $\check{\Lambda}$ follow from arguments along the lines of Spiekerman and Lin (1998). The consistency of $\check{\theta}$ is proved in the Appendix. Also following Spiekerman and Lin (1998), we can show that $\sqrt{K}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges to a mean zero normal distribution and that $\sqrt{K}(\check{\Lambda} - \Lambda_0)$ converges to a mean zero Gaussian process.

Theorem 3. *Under regularity conditions C.1-C.7 in the Appendix, $\sqrt{K}(\check{\theta} - \theta_0)$ converges to a normal distribution with mean zero and variance*

$$\frac{Var(\Xi_1)}{W(\theta_0)^2}.$$

The proof of this theorem and the precise definition of Ξ_1 and $W(\theta_0)$, together with their estimators, can be found in the Appendix.

4 Copula likelihood expression for distributions from the PVF family

The power variance function family of distributions, denoted PVF(α, δ, γ), is a large class of distributions for which Hougaard (2000) states that the Laplace transforms correspond to

$$\mathcal{L}(s) = \exp\left[-\frac{\delta}{\alpha}\left((\gamma+s)^{\alpha} - \gamma^{\alpha}\right)\right]$$

with derivatives

$$\mathcal{L}^{(k)}(s) = (-1)^k \mathcal{L}(s) \sum_{j=1}^{k} c_{k,j}(\alpha)\delta^j(\gamma+s)^{j\alpha-k},$$

where the coefficients $c_{k,j}(\alpha)$ are polynomials of order $k-j$ in α, given by the recursive formula

$$c_{k,1}(\alpha) = \frac{\Gamma(k-\alpha)}{\Gamma(1-\alpha)}, \quad c_{k,k} = 1$$

$$c_{k,j}(\alpha) = c_{k-1,j-1}(\alpha) + c_{k-1,j}(\alpha)(k-1-j\alpha)$$

This allows for a closed form expression of the copula likelihood (3).

Example 1: The one-parameter gamma distribution with density

$$g_\theta(x) = \frac{x^{1/\theta-1}e^{-x/\theta}}{\theta^{1/\theta}\Gamma(1/\theta)}, \quad \theta > 0.$$

is found as the limiting case $\alpha = 0, \delta = \gamma = 1/\theta$. Failure times are independent when θ approaches zero. The Laplace transform is

$$\mathcal{L}(s) = \varphi_\theta(s) = (1+\theta s)^{-1/\theta}$$

which is the generator of the Clayton copula. The Clayton copula has lower tail dependence, which, in a survival context, corresponds to a higher association later in time.

Example 2: The choice $\alpha = \theta, \delta = \theta, \gamma = 0$ leads to the positive stable distribution with density

$$g_\theta(x) = -\frac{1}{\pi x}\sum_{k=1}^{\infty}\frac{\Gamma(k\theta+1)}{k!}(-x^{-\theta})^k \sin(\theta k\pi)$$

with $0 < \theta < 1$. Feller (1971) shows that this density function can be found by Fourier inversion of the Laplace transform

$$\mathcal{L}(s) = \varphi_\theta(s) = e^{-s^\theta}$$

which is the generator of the Gumbel-Hougaard copula. Small values of θ provide large correlation and survival times are independent as θ approaches 1. The Gumbel-Hougaard copula has upper tail dependence, implying a stronger correlation between the lower survival times.

Example 3: Another PVF distribution is obtained by choosing $\alpha = 1/2, \delta = (2\theta)^{-1/2}, \gamma = (2\theta)^{-1}$. This is the inverse Gaussian distribution with variance θ. The density is defined by

$$f_\theta(x) = \sqrt{\frac{1}{2\pi\theta}} x^{-3/2} \exp\left(\frac{-1}{2x\theta}(x-1)^2\right)$$

with $\theta > 0$. The Laplace transform is

$$\mathcal{L}(s) = \varphi_\theta(s) = \exp\left(\frac{1}{\theta} - \left(\frac{1}{\theta^2} + 2\frac{s}{\theta}\right)^{1/2}\right).$$

5 Simulation study

We generate 1000 data sets with 50, 200 or 500 clusters of size varying uniformly between 2 and 50. Survival times are simulated from respectively a Clayton copula with $\theta_0 = 0.2, 0.5, 1.0, 1.5$ or from a Gumbel-Hougaard copula with $\theta_0 = 0.2, 0.5, 0.65, 0.8$, and with, in both settings, Weibull marginal survival functions $S(t) = \lambda t^\rho \exp(\beta' Z)$, choosing $\rho = 1.5$, $\lambda = 0.0316$ and Z a dichotomous covariate with effect $\beta = 3$. The values of the association parameter θ for both copula models are chosen such that the according values of Kendall's tau are comparable. Data are generated using the sampling algorithm of Marshall and Olkin (1988). The censoring distribution is also Weibull, with parameters $(\lambda_C = 0.0274, \rho_C = 1.5)$ and $(\lambda_C = 0.1464, \rho_C = 1.5)$ yielding censoring percentages of 25% and 50%, respectively. The performances of one-stage parametric estimation, two-stage parametric estimation and two-stage semi-parametric estimation are summarized in Tables 2, 3 and 4. For each copula, simulation results are listed in increasing order of association. For the Clayton copula, higher values of θ correspond to a higher degree of association via $\tau = \frac{\theta}{\theta+2}$ whereas the inverse link holds for the Gumbel-Hougaard copula ($\tau = 1 - \theta$). For each degree of association, we report the mean estimated values of $\hat{\theta}$, $\overline{\theta}$ and $\breve{\theta}$ in the first row. Mean standard errors together with the coverage are reported in the second row. Standard errors of one-stage parametric estimators are calculated from the inverse Hessian matrix. In the two-stage parametric approach, standard errors are found via formula (6). In the two-stage semiparametric case, we used the grouped jackknife to obtain standard errors (Lipsitz et al., 1994; Lipsitz and Parzen, 1996). As in the work of Othus and Li (2010) we noted that the variance expression in the two-stage semiparametric estimation method is rather complicated to implement. We assessed the performance of the jackknife procedure in the two-stage parametric model by comparing the standard error through the theoretical expression with a jackknife alternative. Since the results were virtually the same, we only show the standard error calculated from the theoretical expression.

Note that, as the number of clusters increases from $K = 50$ (Table 2) to $K = 200$ (Table 3), standard errors are halved since they are proportional to $1/\sqrt{K}$. For the Gumbel-Hougaard copula, the bias of the estimates are not noticeably affected by an increasing percentage of

censoring, Only when we go from the one-stage parametric estimation method to the two-stage estimation methods we have an increase in the bias. However the standard errors become a bit larger when more censoring is present. For the Clayton copula, we observe that the bias of the estimators increases more when the percentage of censoring increases than in the case of the Gumbel-Hougaard copula. For the standard errors, we see in the Clayton copula similar results as for the Gumbel-Hougaard copula. The combined effect of the increased bias and slightly different standard errors for the Clayton model in comparison of the Gumbel-Hougaard model explain why the coverages are smaller in the Clayton model than in the Gumbel-Hougaard model. A general observation is that biases and standard errors tend to shrink as θ_0 approaches independence. In each of Tables 2, 3 and 4, the largest biases are found in the semiparametric cases where θ_0 has moved far away from independence. The transition from $K = 50$ to $K = 200$ and $K = 500$ leads to a reduction of the bias, which also follows from the asymptotic proofs in the Appendix. However, when the number of clusters is small and the variability of cluster sizes is large, the two-stage parametric and semi-parametric procedures are not recommended. Although computationally more demanding, the one-stage parametric procedure yields the best results in every setting.

6 Modelling time to first insemination in cows clustered in herds

In dairy cattle, the calving interval (the time between two calvings) should be optimally between 12 and 13 months. One of the main factors determining the length of the calving interval is the time from parturition to the time of first insemination (Duchateau and Janssen, 2004). The objective of this study, amongst others, was to quantify the correlation between insemination times of cows within a herd. Insemination at a dairy farm is typically done by the farmer itself, relying on his experience. In this way, we get some insight into this process. The data set includes 181 clusters (farms) of different sizes, ranging from 1 cow to 174 cows. The censoring percentage is 5.5%. The parity of the cow (0 if multiparous, 1 if primiparous) is added as a covariate. In the parametric approach, we first assume a Weibull distribution for the times to first insemination

$$S(t) = \exp(-\lambda \exp(\beta' Z) t^{\rho})$$

and model the association structure by a Clayton copula and a Gumbel-Hougaard copula. In Table 1, the results are listed for the parity effect and association parameter, using the one-stage parametric, two-stage parametric and two-stage semiparametric estimation procedures. In addition, a model with piecewise constant baseline hazard was also fitted, because it has the advantage of a flexible baseline hazard - making it a good alternative for the semiparametric model - but is also parametric , and thus the one-stage estimation procedure can be used. Hereby cutpoints are chosen such that each time interval contains 5% of the events.

In both copula models, the results for the parity effect are similar for all estimation ap-

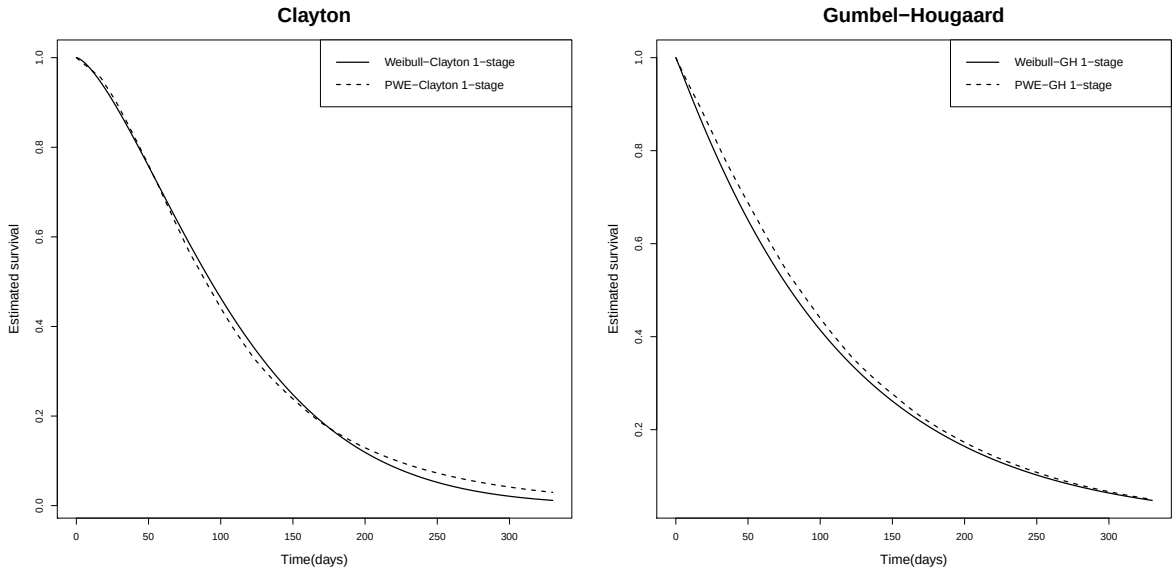

Figure 1: Estimated survival curves for multiparous cows

proaches (see Table 1). The hazard ratio in the one-stage Weibull-Clayton model equals 0.92 (95% CI: $[0.89, 0.95]$), and is 0.95 (95% CI: $[0.92, 0.97]$) for the Weibull-Gumbel-Hougaard model. Both the parametric Weibull and semiparametric two-stage approaches lead to a hazard ratio of 0.94 (95% CI: $[0.90, 0.98]$). For the PWE-Clayton and PWE-Gumbel-Hougaard models, hazard ratios are 0.93 (95% CI: $[0.90, 0.96]$) and 0.94 (95% CI: $[0.92, 0.97]$), respectively. Within each copula model, the parameter estimates for θ vary over the different estimation techniques. The lowest values of θ are observed for the one-stage Weibull models and the highest for the two-stage semiparametric models. Regarding the simulation results in Section 5, we emphasize that the one-stage parametric procedure is most reliable for relatively small sample sizes. If the Weibull assumption is questionable, a piecewise exponential model for the hazard function is recommended.

	Clayton copula				Gumbel-Hougaard copula			
	Weibull one-stage	Weibull two-stage	PWE one-stage	Semipar. two-stage	Weibull one-stage	Weibull two-stage	PWE one-stage	Semipar. two-stage
β	-0.082	-0.066	-0.070	-0.060	-0.055	-0.066	-0.058	-0.060
	(0.017)	(0.022)	(0.016)	(0.021)	(0.013)	(0.022)	(0.014)	(0.021)
θ	0.212	0.324	0.352	0.447	0.624	0.766	0.661	0.790
	(0.015)	(0.050)	(0.034)	(0.063)	(0.016)	(0.018)	(0.013)	(0.016)

Table 1: Estimation results for time to first insemination data

A visual check of the estimated marginal survival curves (see Figure 1) reveals why the difference between the estimated association parameter θ in the one-stage Weibull-Clayton and PWE-

Clayton is so large (0.212 versus 0.352). The difference between the estimated marginal survival functions is largest for later times, which are the times when the Clayton copula imposes a higher dependency. If the Weibull assumption is incorrect, the estimated association parameter will also lack accuracy. In this example, we used both a Clayton and a Gumbel-Hougaard copula to illustrate our techniques. At this moment, we did not focus on a goodness-of-fit test for the selection of the copula function. This will be done in the future.

7 Discussion

The current copula methodology only allows the modelling of multivariate survival data that are grouped in clusters of small and equal size. A new formulation for the likelihood of Archimedean copula models for survival data is developed, that allows for clusters of large and variable size. The failure times within a cluster are assumed to be exchangeable and the whole data set is used to estimate a common marginal baseline survival. The survival functions of subjects differ through the incorporation of covariates (possibly time-dependent). For copula members of the PVF family, a closed form expression of the likelihood exists, whereas other choices require numerical integration. We investigated the parametric one-stage and two-stage approach as well as the semiparametric two-stage approach and derived asymptotic results for the estimators under a reasonable set of conditions. Simulation results show that all three methods work well for cluster sizes ranging from 2 to 50. Even larger clusters can be attained, at the cost of larger computing time. For samples with less than 100 clusters, the two-stage estimation approaches are not recommended since they lead to larger bias and less coverage. As an alternative to the flexible semiparametric model, a piecewise constant hazard (or, by extension, e.g. splines) can be used while modelling the marginal survival function. This article is an extension of the work of Shih and Louis (1995), who derived founding results for bivariate data, and the work of Glidden (2000), who investigated the two-stage semiparametric model for the Clayton copula, as it describes the use of copula functions for clusters with large and varying cluster size.

Acknowledgements

The authors would like to gratefully acknowledge the financial support from the Interuniversity Attraction Poles Programme (IAP-network P7/06) of the Belgian Science Policy Office. For the simulations we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI. Furthermore we would like to thank the editor, associate editor and two referees for their comments on the first version of this manuscript. This helped us to improve it and to clarify our research.

Copula model	τ	θ_0	0% censoring Parametric one-stage	0% censoring Parametric two-stage	0% censoring Semiparametric two-stage	25% censoring Parametric one-stage	25% censoring Parametric two-stage	25% censoring Semiparametric two-stage	50% censoring Parametric one-stage	50% censoring Parametric two-stage	50% censoring Semiparametric two-stage
Clayton	0.09	0.2	0.197	0.193	0.191	0.196	0.194	0.194	0.197	0.195	0.195
			(0.043; 93.1%)	(0.042; 89.7%)	(0.045; 85.9%)	(0.047; 92.0%)	(0.047; 90.7%)	(0.049; 88.9%)	(0.055; 92.1%)	(0.055; 91.3%)	(0.056, 88.9%)
	0.2	0.5	0.498	0.486	0.463	0.496	0.489	0.479	0.495	0.491	0.485
			(0.084; 93.2%)	(0.091; 84.3%)	(0.010; 76.8%)	(0.091; 92.9%)	(0.097; 88.1%)	(0.105; 85.3%)	(0.101; 92.7%)	(0.106; 89.8%)	(0.113; 87.8%)
	0.33	1.0	0.997	0.973	0.875	0.996	0.981	0.938	0.997	0.990	0.959
			(0.160; 93.5%)	(0.176; 81.9%)	(0.174; 71.8%)	(0.166; 92.9%)	(0.182; 86.9%)	(0.195; 81.7%)	(0.178; 92.3%)	(0.194; 88.9%)	(0.205; 85.7%)
	0.43	1.5	1.479	1.436	1.226	1.478	1.451	1.365	1.476	1.469	1.402
			(0.234; 92.1%)	(0.253; 83.9%)	(0.229; 63.0%)	(0.240; 92.6%)	(0.262; 87.5%)	(0.273; 81.3%)	(0.252; 91.7%)	(0.278; 88.5%)	(0.287; 84.9%)
G-H	0.2	0.8	0.803	0.801	0.803	0.804	0.802	0.803	0.804	0.802	0.804
			(0.034; 93.6%)	(0.042; 88.6%)	(0.041; 89.0%)	(0.036; 94.3%)	(0.045; 89.0%)	(0.044; 87.7%)	(0.039; 94.9%)	(0.048; 87.8%)	(0.048; 86.0%)
	0.35	0.65	0.656	0.655	0.661	0.656	0.656	0.662	0.656	0.656	0.664
			(0.040; 93.5%)	(0.048; 89.4%)	(0.049; 89.5%)	(0.042; 93.3%)	(0.051; 89.2%)	(0.052; 88.9%)	(0.045; 94.6%)	(0.055; 88.0%)	(0.056; 86.4%)
	0.5	0.5	0.507	0.507	0.521	0.508	0.508	0.522	0.507	0.509	0.525
			(0.040; 93.3%)	(0.046; 91.2%)	(0.047; 90.4%)	(0.041; 93.6%)	(0.048; 90.5%)	(0.050; 90.2%)	(0.043; 94.3%)	(0.051; 88.4%)	(0.054; 86.9%)
	0.8	0.2	0.205	0.208	0.247	0.205	0.209	0.250	0.205	0.211	0.258
			(0.022; 94.7%)	(0.023; 92.3%)	(0.030; 68.5%)	(0.022; 94.2%)	(0.025; 92.6%)	(0.032; 67.2%)	(0.023; 95.1%)	(0.026; 89.7%)	(0.035; 60.9%)

Table 2: Simulation results for 50 clusters of varying sizes ranging from 2 to 50

Copula model	τ	θ_0	0% censoring Parametric one-stage	0% censoring Parametric two-stage	0% censoring Semiparametric two-stage	25% censoring Parametric one-stage	25% censoring Parametric two-stage	25% censoring Semiparametric two-stage	50% censoring Parametric one-stage	50% censoring Parametric two-stage	50% censoring Semiparametric two-stage
Clayton	0.09	0.2	0.199	0.198	0.197	0.199	0.198	0.198	0.199	0.199	0.199
			(0.021; 94.4%)	(0.022; 93.3%)	(0.025; 90.8%)	(0.024; 94.4%)	(0.024; 94.0%)	(0.025; 92.6%)	(0.027; 95.3%)	(0.028; 94.8%)	(0.029; 94.0%)
	0.2	0.5	0.498	0.498	0.489	0.498	0.498	0.495	0.498	0.499	0.497
			(0.042; 94.3%)	(0.052; 90.8%)	(0.059; 88.8%)	(0.045; 94.6%)	(0.052; 92.6%)	(0.057; 92.0%)	(0.050; 93.3%)	(0.055; 93.4%)	(0.060; 92.5%)
	0.33	1.0	0.994	0.990	0.953	0.993	0.990	0.978	0.994	0.992	0.984
			(0.079; 95.3%)	(0.101; 90.8%)	(0.110; 86.8%)	(0.083; 94.3%)	(0.099; 92.6%)	(0.108; 90.7%)	(0.088; 95.1%)	(0.102; 93.4%)	(0.109; 92.0%)
	0.43	1.5	1.494	1.484	1.401	1.494	1.488	1.463	1.496	1.491	1.472
			(0.118; 94.2%)	(0.147; 90.1%)	(0.152; 82.2%)	(0.121; 94.4%)	(0.145; 91.2%)	(0.155; 90.4%)	(0.127; 94.8%)	(0.148; 91.8%)	(0.157; 91.1%)
G-H	0.2	0.8	0.802	0.801	0.802	0.801	0.801	0.801	0.801	0.800	0.801
			(0.017; 95.8%)	(0.022; 93.2%)	(0.022; 92.4%)	(0.018; 94.6%)	(0.024; 92.4%)	(0.024; 91.6%)	(0.020; 95.5%)	(0.026; 92.2%)	(0.026; 91.6%)
	0.35	0.65	0.652	0.652	0.654	0.652	0.652	0.654	0.652	0.652	0.655
			(0.020; 95.2%)	(0.025; 93.9%)	(0.026; 93.0%)	(0.021; 95.0%)	(0.027; 93.3%)	(0.028; 93.4%)	(0.022; 95.4%)	(0.030; 92.8%)	(0.030; 93.0%)
	0.5	0.5	0.503	0.503	0.507	0.502	0.503	0.508	0.502	0.503	0.509
			(0.020; 94.8%)	(0.024; 93.8%)	(0.024; 93.4%)	(0.020; 94.7%)	(0.025; 93.7%)	(0.026; 93.3%)	(0.021; 95.1%)	(0.028; 93.2%)	(0.029; 93.0%)
	0.8	0.2	0.201	0.202	0.215	0.201	0.203	0.217	0.201	0.203	0.220
			(0.011; 95.3%)	(0.012; 94.7%)	(0.014; 81.4%)	(0.011; 94.0%)	(0.013; 94.5%)	(0.015; 81.2%)	(0.011; 94.7%)	(0.013; 93.9%)	(0.016; 76.9%)

Table 3: Simulation results for 200 clusters of varying sizes ranging from 2 to 50

Copula model	τ	θ_0	0% censoring Parametric one-stage	0% censoring Parametric two-stage	0% censoring Semiparametric two-stage	25% censoring Parametric one-stage	25% censoring Parametric two-stage	25% censoring Semiparametric two-stage	50% censoring Parametric one-stage	50% censoring Parametric two-stage	50% censoring Semiparametric two-stage
Clayton	0.09	0.2	0.200	0.200	0.199	0.200	0.200	0.200	0.200	0.199	0.200
			(0.013; 95.4%)	(0.014; 94.2%)	(0.016; 92.6%)	(0.015; 95.3%)	(0.016; 94.1%)	(0.017; 93.7%)	(0.017; 94.9%)	(0.018; 95.2%)	(0.019; 94.8%)
	0.2	0.5	0.501	0.499	0.493	0.501	0.499	0.498	0.501	0.500	0.499
			(0.027; 95.4%)	(0.033; 92.2%)	(0.039; 89.9%)	(0.029; 94.9%)	(0.033; 92.8%)	(0.037; 91.7%)	(0.032; 94.6%)	(0.035; 93.4%)	(0.038; 93.2%)
	0.33	1.0	0.999	0.994	0.973	1.000	0.996	0.990	0.999	0.997	0.992
			(0.050; 94.8%)	(0.065; 91.8%)	(0.072; 89.1%)	(0.053; 94.1%)	(0.064; 92.9%)	(0.070; 92.1%)	(0.056; 94.2%)	(0.065; 93.0%)	(0.070; 93.2%)
	0.43	1.5	1.498	1.496	1.453	1.498	1.497	1.485	1.497	1.498	1.490
			(0.075; 93.8%)	(0.098; 93.4%)	(0.104; 88.8%)	(0.077; 94.3%)	(0.095; 93.7%)	(0.101; 93.4%)	(0.081; 93.5%)	(0.095; 93.4%)	(0.102; 93.9%)
G-H	0.2	0.8	0.800	0.801	0.801	0.801	0.801	0.801	0.801	0.801	0.802
			(0.011; 93.6%)	(0.014; 94.3%)	(0.014; 92.9%)	(0.011; 95.5%)	(0.015; 93.3%)	(0.015; 93.0%)	(0.013; 95.4%)	(0.017; 93.6%)	(0.017; 93.0%)
	0.35	0.65	0.651	0.652	0.653	0.651	0.652	0.653	0.652	0.652	0.654
			(0.013; 95.4%)	(0.016; 95.3%)	(0.017; 95.1%)	(0.013; 95.9%)	(0.017; 93.7%)	(0.018; 93.6%)	(0.014; 94.6%)	(0.019; 93.8%)	(0.020; 92.9%)
	0.5	0.5	0.501	0.502	0.504	0.501	0.502	0.505	0.502	0.502	0.505
			(0.013; 96.9%)	(0.015; 95.0%)	(0.016; 94.8%)	(0.013; 94.9%)	(0.016; 93.8%)	(0.017; 93.4%)	(0.014; 95.8%)	(0.018; 93.9%)	(0.018; 93.6%)
	0.8	0.2	0.201	0.201	0.208	0.201	0.201	0.209	0.201	0.202	0.211
			(0.007; 95.7%)	(0.008; 95.4%)	(0.009; 86.7%)	(0.007; 95.9%)	(0.008; 95.4%)	(0.009; 86.1%)	(0.007; 95.0%)	(0.009; 94.6%)	(0.010; 83.2%)

Table 4: Simulation results for 500 clusters of varying sizes ranging from 2 to 50

References

Andersen, E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, 11:333–350.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220.

Cox, D. R. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall.

Duchateau, L. and Janssen, P. (2004). Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. *Biometrics*, 60(3):608–614.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. Wiley.

Glidden, D. V. (2000). A two-stage estimator of the dependence parameter for the clayton-oakes model. *Lifetime Data Analysis*, 6:141–156.

Goethals, K., Janssen, P., and Duchateau, L. (2008). Frailty models and copulas: similarities and differences. *Journal of Applied Statistics*, 35(9):1071 –1079.

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

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer.

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

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Lipsitz, S. R., Dear, K. B., and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, 50:842–846.

Lipsitz, S. R. and Parzen, M. (1996). A jackknife estimator of variance for cox regression for correlated survival data. *Biometrics*, 52:291–298.

Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, 83(403):834 – 841.

Massonnet, G., Janssen, P., and Duchateau, L. (2009). Modelling udder infection data using copula models for quadruples. *Journal of Statistical Planning and Inference*, 139:3865 –3877.

Nelsen, R. B. (2006). *An Introduction to copulas*. Springer.

Othus, M. and Li, Y. (2010). A gaussian copula model for multivariate survival data. *Statistics in Biosciences*, 2:154–179.

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

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93:1164–1175.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall.

Xu, J. (1996). Statistical modelling and inference for multivariate and longitudinal discrete response data. *Ph.D. Thesis*.

Appendix: Theorems and proofs

Proof of Theorem 1. Let $\boldsymbol{\beta}_0$ denote the true parameter vector for the margins. Expanding the score function $\mathbf{U}^*_{\boldsymbol{\beta}}$ in a Taylor series around $\boldsymbol{\beta}_0$ and evaluating it at $\boldsymbol{\beta} = \overline{\boldsymbol{\beta}}$, we get under regularity conditions of maximum likelihood theory

$$\mathbf{U}^*_{\boldsymbol{\beta}}(\overline{\boldsymbol{\beta}}) = \mathbf{0} = \mathbf{U}^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \left.\frac{\partial \mathbf{U}^*_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(\sqrt{K}).$$

Similarly,

$$U_\theta(\overline{\boldsymbol{\beta}}, \overline{\theta}) = 0 = U_\theta(\boldsymbol{\beta}_0, \theta_0) + \left.\frac{\partial U_\theta}{\partial \boldsymbol{\beta}}\right|_{(\boldsymbol{\beta},\theta)=(\boldsymbol{\beta}_0,\theta_0)} (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left.\frac{\partial U_\theta}{\partial \theta}\right|_{(\boldsymbol{\beta},\theta)=(\boldsymbol{\beta}_0,\theta_0)} (\overline{\theta} - \theta_0) + o_p(\sqrt{K}).$$

By the law of large numbers, as $K \to \infty$,

$$\begin{aligned}
-\frac{1}{K}\left.\frac{\partial \mathbf{U}^*_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &= \frac{1}{K}\sum_{i=1}^{K} -\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{U}^*_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}_0) \to \mathbf{I}^* = E\left[-\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{U}^*_{1,\boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right] \\
-\frac{1}{K}\left.\frac{\partial U_\theta}{\partial \boldsymbol{\beta}}\right|_{(\boldsymbol{\beta},\theta)=(\boldsymbol{\beta}_0,\theta_0)} &= \frac{1}{K}\sum_{i=1}^{K} -\frac{\partial}{\partial \boldsymbol{\beta}} U_{i,\theta}(\boldsymbol{\beta}_0, \theta_0) \to \mathbf{I}_{\theta\boldsymbol{\beta}} \\
-\frac{1}{K}\left.\frac{\partial U_\theta}{\partial \theta}\right|_{(\boldsymbol{\beta},\theta)=(\boldsymbol{\beta}_0,\theta_0)} &= \frac{1}{K}\sum_{i=1}^{K} -\frac{\partial}{\partial \theta} U_{i,\theta}(\boldsymbol{\beta}_0, \theta_0) \to I_{\theta\theta}.
\end{aligned}$$

Hence

$$\frac{1}{\sqrt{K}} \begin{pmatrix} \mathbf{U}^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0) \\ U_\theta(\boldsymbol{\beta}_0, \theta_0) \end{pmatrix} \to \sqrt{K} \begin{pmatrix} \mathbf{I}^* & 0 \\ \mathbf{I}_{\theta\boldsymbol{\beta}} & I_{\theta\theta} \end{pmatrix} \begin{pmatrix} \overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \overline{\theta} - \theta_0 \end{pmatrix}.$$

By the central limit theorem, $\frac{1}{\sqrt{K}} \begin{pmatrix} \mathbf{U}^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0) \\ U_\theta(\boldsymbol{\beta}_0, \theta_0) \end{pmatrix}$ converges to multivariate normal with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance-covariance matrix $\begin{pmatrix} \mathbf{V} & 0 \\ 0 & I_{\theta\theta} \end{pmatrix}$ with $\mathbf{V} = \text{Var}\left(\mathbf{U}^*_{1,\boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right) = E\left[\mathbf{U}^*_{1,\boldsymbol{\beta}}(\boldsymbol{\beta}_0)^2\right]$. Thus, $\sqrt{K}\begin{pmatrix} \overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \overline{\theta} - \theta_0 \end{pmatrix}$ converges to multivariate normal with mean vector zero and variance-covariance matrix

$$\begin{pmatrix} \mathbf{I}^* & \mathbf{0} \\ \mathbf{I}_{\theta\boldsymbol{\beta}} & I_{\theta\theta} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & I_{\theta\theta} \end{pmatrix} \begin{pmatrix} \mathbf{I}^* & \mathbf{0} \\ \mathbf{I}_{\theta\boldsymbol{\beta}} & I_{\theta\theta} \end{pmatrix}^{-1^T} = \begin{pmatrix} (\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1^T} & \frac{-(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1^T}\mathbf{I}_{\boldsymbol{\beta}\theta}}{I_{\theta\theta}} \\ \frac{-\mathbf{I}_{\theta\boldsymbol{\beta}}(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1^T}}{I_{\theta\theta}} & \frac{1}{I_{\theta\theta}} + \frac{\mathbf{I}_{\theta\boldsymbol{\beta}}(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1^T}\mathbf{I}_{\boldsymbol{\beta}\theta}}{I^2_{\theta\theta}} \end{pmatrix}.$$

The lower right element of this matrix is the asymptotic variance of $\sqrt{K}(\overline{\theta} - \theta_0)$ and we denote this by σ^2.

$$\sigma^2 = \frac{1}{I_{\theta\theta}} + \frac{\mathbf{I}_{\theta\boldsymbol{\beta}}(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}\mathbf{I}_{\boldsymbol{\beta}\theta}}{I^2_{\theta\theta}}.$$

Before we prove Theorem 2 and 3, we first introduce some notation.

$$
\begin{aligned}
Y_{ij}(t) &= I_{\{X_{ij} \geq t\}} \\
\check{\Lambda}(t) &= \int_0^t \frac{d \sum_{i=1}^K \sum_{j=1}^{n_i} \delta_{ij} I_{\{X_{ij} \leq u\}}}{\sum_{i=1}^K \sum_{j=1}^{n_i} Y_{ij}(u) \exp[\check{\boldsymbol{\beta}}' \boldsymbol{Z}_{ij}(u)]} = \sum_{i=1}^K \sum_{j=1}^{n_i} \frac{\delta_{ij} I_{\{X_{ij} \leq t\}}}{\sum_{k=1}^K \sum_{l=1}^{n_k} I_{\{X_{kl} \leq X_{ij}\}} \exp[\check{\boldsymbol{\beta}}' \boldsymbol{Z}_{kl}(X_{ij})]} \\
H_{ij} &= \exp\left(- \int_0^\tau Y_{ij}(u) \exp[\boldsymbol{\beta}' \boldsymbol{Z}_{ij}(u)] d\Lambda(u) \right) \\
H_{ij}^0 &= \exp\left(- \int_0^\tau Y_{ij}(u) \exp[\boldsymbol{\beta}_0' \boldsymbol{Z}_{ij}(u)] d\Lambda_0(u) \right) \\
\check{H}_{ij} &= \exp\left(- \int_0^\tau Y_{ij}(u) \exp[\check{\boldsymbol{\beta}}' \boldsymbol{Z}_{ij}(u)] d\check{\Lambda}(u) \right) \\
H_{ij}(t) &= \exp\left(- \int_0^\tau Y_{ij}(u) \exp[\boldsymbol{\beta}' \boldsymbol{Z}_{ij}(u)] d(\Lambda + t(\Gamma - \Lambda))(u) \right)
\end{aligned}
$$

Note that $H_{ij} = H_{ij}(0)$.

$$
\begin{aligned}
L(\theta; \boldsymbol{\beta}, \Lambda) &= \prod_{i=1}^K L_i(\theta; \boldsymbol{\beta}, \Lambda) \\
&= \prod_{i=1}^K \left(\prod_{j=1}^{n_i} \left[\frac{1}{\varphi_\theta' \left(\varphi_\theta^{-1}(H_{ij}) \right)} \right]^{\delta_{ij}} \right) \varphi_\theta^{(d_i)} \left(\sum_{j=1}^{n_i} \varphi_\theta^{-1}(H_{ij}) \right) \\
l_K(\theta) &= K^{-1} \log L(\theta; \boldsymbol{\beta}, \Lambda) \\
&= K^{-1} \sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} \delta_{ij} \log \left[\frac{1}{\varphi_\theta' \left(\varphi_\theta^{-1}(H_{ij}) \right)} \right] + \log \varphi_\theta^{(d_i)} \left(\sum_{j=1}^{n_i} \varphi_\theta^{-1}(H_{ij}) \right) \right\} \\
l_{K0}(\theta) &= K^{-1} \log L(\theta; \boldsymbol{\beta}_0, \Lambda_0) \\
\check{l}_K(\theta) &= K^{-1} \log L(\theta; \check{\boldsymbol{\beta}}, \check{\Lambda}) \\
U_K(\theta) &= \frac{\partial l_K(\theta)}{\partial \theta} = K^{-1} \frac{\partial \log L(\theta; \boldsymbol{\beta}, \Lambda)}{\partial \theta} \\
&= K^{-1} \sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} \delta_{ij} \left[\varphi_\theta' \left(\varphi_\theta^{-1}(H_{ij}) \right) \right] \frac{\partial}{\partial \theta} \left[\varphi_\theta' \left(\varphi_\theta^{-1}(H_{ij}) \right) \right]^{-1} \right. \\
&\qquad \left. + \left[\varphi_\theta^{(d_i)} \left(\sum_{j=1}^{n_i} \varphi_\theta^{-1}(H_{ij}) \right) \right]^{-1} \frac{\partial}{\partial \theta} \left[\varphi_\theta^{(d_i)} \left(\sum_{j=1}^{n_i} \varphi_\theta^{-1}(H_{ij}) \right) \right] \right\} \\
U_{K0}(\theta) &= \frac{\partial l_{K0}(\theta)}{\partial \theta} = K^{-1} \frac{\partial \log L(\theta; \boldsymbol{\beta}_0, \Lambda_0)}{\partial \theta} \\
\check{U}_K(\theta) &= \frac{\partial \check{l}_K(\theta)}{\partial \theta} = K^{-1} \frac{\partial \log L(\theta; \check{\boldsymbol{\beta}}, \check{\Lambda})}{\partial \theta}
\end{aligned}
$$

We copy the following notation from Spiekerman and Lin (1998) where $\boldsymbol{a}^{\otimes 0} = 1, \boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}' \boldsymbol{a}$:

$$
\begin{aligned}
&\boldsymbol{S}^{(r)}(\boldsymbol{\beta}, t) = K^{-1} \sum_{i=1}^K \sum_{j=1}^{n_i} Y_{ij}(t) \exp[\boldsymbol{\beta}' \boldsymbol{Z}_{ij}(t)] \boldsymbol{Z}_{ij}(t)^{\otimes r}, \qquad \boldsymbol{s}^{(r)} = E\left[\boldsymbol{S}^{(r)}(\boldsymbol{\beta}, t) \right] \quad (r = 0, 1, 2) \\
&\boldsymbol{E}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}, \qquad \boldsymbol{e}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \\
&\boldsymbol{V}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{S}^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \boldsymbol{E}(\boldsymbol{\beta}, t)^{\otimes 2}, \qquad \boldsymbol{v}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \boldsymbol{e}(\boldsymbol{\beta}, t)^{\otimes 2}
\end{aligned}
$$

Assume the following regularity conditions where $\tau > 0$ is a constant (e.g. end of study time).

C1. $\boldsymbol{\beta}$ is in a compact subset of $\mathbb{R}^p$

C2. $\Lambda(\tau) < \infty$

C3. $\theta \in \nu$, where ν is a compact subset of Θ

C4. $P(C_{ij} \geq t \quad \forall t \in [0,\tau]) > \delta_c > 0$ for $i = 1,\ldots,K$ and $j = 1,\ldots,n_i$

C5. Write $\boldsymbol{Z}_{ij}(t) = \{Z_{ij1}(t),\ldots,Z_{ijp}(t)\}$. For $i = 1,\ldots,K, j = 1\ldots,n_i, k = 1,\ldots,p$

$$|Z_{ijk}(0)| + \int_0^\tau |dZ_{ijk}(t)| \leq B_Z < \infty \quad \text{a.s. for some constant } B_Z$$

C6. $E\left[\log \frac{L_i(\theta_1;\boldsymbol{\beta},\Lambda)}{L_i(\theta_2;\boldsymbol{\beta},\Lambda)}\right]$ exists for all $\theta_1,\theta_2 \in \Theta, i = 1,\ldots,K$

C7. $\boldsymbol{A} = \int_0^\tau \boldsymbol{v}(\boldsymbol{\beta}_0,u)s^{(0)}(\boldsymbol{\beta}_0,u)d\Lambda_0(u)$ is positive definite.

Proof of Theorem 2. The results for $\check{\boldsymbol{\beta}}$ and $\check{\Lambda}$ follow from arguments along the lines of Spiekerman and Lin (1998). We will now show the consistency of $\check{\theta}$ using ideas of Othus and Li (2010).

To account for the fact that plug-in estimates of $\boldsymbol{\beta}$ and Λ are used in the likelihood for θ, we will need to take a Taylor series expansion of the likelihood of θ around $\boldsymbol{\beta}_0$ and Λ_0. Since Λ_0 is an unspecified function, this expansion will need to include a functional expansion term. An expansion using Hadamard derivatives is appropriate for this situation. Hereto, we must verify that the log-likelihood $l_K(\theta)$ is Hadamard differentiable with respect to Λ.

We find the Hadamard derivative of l_K w.r.t. Λ at $\Gamma - \Lambda \in BV[0,\tau]$ by taking the derivative of $K^{-1}\log L(\theta;\boldsymbol{\beta},\Lambda + t(\Gamma-\Lambda))$ with respect to t en then putting $t = 0$:

$$\frac{d}{dt}\left[K^{-1}\log L(\theta;\boldsymbol{\beta},\Lambda + t(\Gamma-\Lambda))\right]\Big|_{t=0} = \int_0^\tau \zeta_K(\theta;\Lambda)(u)d(\Gamma-\Lambda)(u)$$

where

$$\zeta_K(\theta;\Lambda)(u) = K^{-1}\sum_{i=1}^{K}\sum_{j=1}^{n_i} D_{ij}^l Y_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}(u)]$$

and

$$D_{ij}^l = \left\{\delta_{ij}\frac{-\varphi_\theta''(\varphi_\theta^{-1}(H_{ij}))}{\varphi_\theta'(\varphi_\theta^{-1}(H_{ij}))} + \frac{\varphi_\theta^{(d_i+1)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)}{\varphi_\theta^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)}\right\}\frac{-H_{ij}}{\varphi_\theta'(\varphi_\theta^{-1}(H_{ij}))}.$$

The derivative of $l_K(\theta)$ w.r.t. $\boldsymbol{\beta}$ is

$$\zeta_K(\theta;\boldsymbol{\beta}) = K^{-1}\sum_{i=1}^{K}\sum_{j=1}^{n_i} D_{ij}^l\left(\int_0^\tau Y_{ij}(u)\boldsymbol{Z}_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}(u)]d\Lambda(u)\right).$$

To prove consistency for $\check{\theta}$, we will require $||\zeta_K(\theta;\Lambda)||_\infty$ and $||\zeta_K(\theta;\boldsymbol{\beta})||$ to be bounded. This can be obtained when the common factor $||D_{ij}^l||_\infty$ is bounded and also the terms unique to $\zeta_K(\theta;\boldsymbol{\beta})$ and $\zeta_K(\theta;\Lambda)$ have to be bounded. This requirement is not too restrictive, e.g. for the Clayton copula we have

$$||D_{ij}^l||_\infty = \left|\left|\delta_{ij}(1+\theta) - \frac{(1+d_i\theta)H_{ij}^{-\theta}}{\left(-n_i+1+\sum_{j=1}^{n_i}H_{ij}^{-\theta}\right)}\right|\right|_\infty.$$

Due to the definition of H_{ij} and condition C2, this expression is bounded. By condition C5,

$$||Y_{ij}\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}]||_\infty \quad \text{and} \quad \left|\left|\int_0^\tau Y_{ij}(u)\boldsymbol{Z}_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij(u)}d\Lambda(u)]\right|\right| \quad \text{are bounded.}$$

An expansion of $\check{l}_K(\theta)$ around $\boldsymbol{\beta}_0$ and Λ_0 can be written as

$$\check{l}_K(\theta) = l_{K0}(\theta) + \zeta_K(\theta;\boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \int_0^\tau \zeta_K(\theta;\Lambda_0)(t)d(\check{\Lambda}-\Lambda_0)(t) + R.$$

Another (intuitive) notation is:

$$l_{K,\theta}(\check{\boldsymbol{\beta}},\check{\Lambda}) = l_{K,\theta}(\boldsymbol{\beta}_0,\Lambda_0) + \frac{\partial}{\partial\boldsymbol{\beta}} l_{K,\theta}(\boldsymbol{\beta}_0,\Lambda_0)(\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + \frac{\partial}{\partial\Lambda} l_{K,\theta}(\boldsymbol{\beta}_0,\Lambda_0)(\check{\Lambda}-\Lambda_0) + R.$$

The remainder term R is of order $o_p\left(\max\{||\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||, ||\check{\Lambda}-\Lambda_0||_\infty\}\right)$. This can be seen from the definition of Hadamard differentiability, since

$$\left|\left|\frac{l_{K,\theta}(\boldsymbol{\beta},\Lambda_0+t(\check{\Lambda}-\Lambda_0)) - l_{K,\theta}(\boldsymbol{\beta},\check{\Lambda})}{t} - \frac{\partial}{\partial\Lambda} l_{K,\theta}(\boldsymbol{\beta},\Lambda_0)(\check{\Lambda}-\Lambda_0)\right|\right|_\infty \to 0, \qquad \text{as } t\downarrow 0,$$

uniformly in $\check{\Lambda}-\Lambda_0$ in all compact subsets of $\mathbb{D}$, the space of cumulative hazard functions. Since $\check{\boldsymbol{\beta}}$ is consistent and $\check{\Lambda}$ is uniformly consistent (Spiekerman and Lin, 1998), $R = o_p(1)$.

In order to prove $\check{\theta}$ is consistent we will need to verify the uniform convergence of the log-likelihood with the plug-in estimate of Λ to the expected value of the log-likelihood evaluated at the true value of Λ, denoted $l_{K0}(\theta)$:

$$\sup_{\theta\in\nu} |\check{l}_K(\theta) - E[l_{K0}(\theta)]| = o_p(1). \tag{7}$$

This can be shown as follows:

$$\check{l}_K(\theta) - E[l_{K0}(\theta)] = l_{K0}(\theta) - E[l_{K0}(\theta)] + \zeta_K(\theta;\boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + \int_0^\tau \zeta_K(\theta;\Lambda_0)(t)d(\check{\Lambda}-\Lambda_0)(t) + R.$$

Due to the law of large numbers, for fixed θ,

$$l_{K0}(\theta) - E[l_{K0}(\theta)] \xrightarrow{p} 0. \tag{8}$$

Since $||\zeta_K(\theta;\boldsymbol{\beta})||$ is bounded, say $||\zeta_K(\theta;\boldsymbol{\beta})|| \le M_1$, we have

$$\sup_{\theta\in\nu} \left|\zeta_K(\theta;\boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right| \le M_1||\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||. \tag{9}$$

Since $||\zeta_K(\theta;\Lambda)(u)||_\infty$ is bounded, say $||\zeta_K(\theta;\Lambda)(u)||_\infty \le M_2$, we have

$$\sup_{\theta\in\nu} \left|\int_0^\tau \zeta_K(\theta;\Lambda)(t)d(\check{\Lambda}-\Lambda_0)(t)\right| \le M_2||\check{\Lambda}-\Lambda_0||_\infty. \tag{10}$$

Therefore

$$\sup_{\theta\in\nu} \left|\check{l}_K(\theta) - E[l_{K0}(\theta)]\right| \le \sup_{\theta\in\nu} |l_{K0}(\theta) - E[l_{K0}(\theta)]| + M_1||\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|| + M_2||\check{\Lambda}-\Lambda_0||_\infty + R.$$

Using (8), the consistency of $\check{\boldsymbol{\beta}}$, the uniform consistency of $\check{\Lambda}$ and the fact that $R = o_p(1)$, we get

$$\sup_{\theta\in\nu} \left|\check{l}_K(\theta) - E[l_{K0}(\theta)]\right| = o_p(1).$$

Finally, in order to verify that $\check{\theta}$ is consistent, we will need to show that the expected log-likelihood is maximized at the truth:

$$E[l_{K0}(\theta)] - E[l_{K0}(\theta_0)] < 0. \tag{11}$$

Due to independence between clusters and the fact that all lower dimensional copulas can be regarded as margins of the highest dimensional copula, the log-likelihood $l_K(\theta)$ can be written as a sum of i.i.d. random variables

$$K^{-1}\sum_{i=1}^K \log L_i(\theta;\boldsymbol{\beta},\Lambda)$$

with

$$\begin{aligned} L_i &= (-1)^{d_i}\frac{\partial^{d_i}}{\partial\{\delta_{ij}=1\}} S(y_{i1},\ldots,y_{i,n_i}) \\ &= \left(\prod_{j=1}^{n_i}\left[\frac{1}{\varphi_\theta'\left(\varphi_\theta^{-1}\left(e^{-\Lambda(y_{ij})}\right)\right)}\right]^{\delta_{ij}}\right)\varphi_\theta^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}\left(e^{-\Lambda(y_{ij})}\right)\right) \end{aligned}$$

where $\partial\{\delta_{ij}=1\}$ is the set of uncensored individuals in cluster i.

Take $\theta \neq \theta_0$. The law of large numbers, Jensen's inequality and condition C6 imply that

$$\begin{aligned}
\lim_{K\to\infty} l_{K0}(\theta) - l_{K0}(\theta_0) &= E[l_{K0}(\theta)] - E[l_{K0}(\theta_0)] \\
&= E\left[K^{-1}\sum_{i=1}^{K}\log L_i(\theta;\boldsymbol{\beta}_0,\Lambda_0)\right] - E\left[K^{-1}\sum_{i=1}^{K}\log L_i(\theta_0;\boldsymbol{\beta}_0,\Lambda_0)\right] \\
&= E\left[\log L_1(\theta;\boldsymbol{\beta}_0,\Lambda_0) - \log L_1(\theta_0;\boldsymbol{\beta}_0,\Lambda_0)\right] \\
&= E\left[\log \frac{L_1(\theta;\boldsymbol{\beta}_0,\Lambda_0)}{L_1(\theta_0;\boldsymbol{\beta}_0,\Lambda_0)}\right] \\
&\leq \log E\left[\frac{L_1(\theta;\boldsymbol{\beta}_0,\Lambda_0)}{L_1(\theta_0;\boldsymbol{\beta}_0,\Lambda_0)}\right] \\
&= \log 1 \\
&= 0.
\end{aligned}$$

The before last equality results from $L_1(\theta;\boldsymbol{\beta}_0,\Lambda_0)$ being the contribution of cluster 1 to the likelihood $L(\theta;\boldsymbol{\beta}_0,\Lambda_0)$, which is the joint density function of $(y_{11},\ldots,y_{1,n_1};\delta_{11},\ldots,\delta_{1,n_1})$.
Since $\check{\theta}$ maximizes $\check{l}_K(\theta)$, (7) implies that

$$0 \leq \check{l}_K(\check{\theta}) - \check{l}_K(\theta_0) = \check{l}_K(\check{\theta}) - \check{l}_K(\theta_0) + E[l_{K0}(\theta_0)] - E[l_{K0}(\theta_0)] = \check{l}_K(\check{\theta}) - E[l_{K0}(\theta_0)] + o_p(1)$$

$$\Downarrow$$

$$E[l_{K0}(\theta_0)] \leq \check{l}_K(\check{\theta}) + o_p(1).$$

Subtract $E[l_{K0}(\check{\theta})]$ from each side of the inequality to write

$$E[l_{K0}(\theta_0)] - E[l_{K0}(\check{\theta})] \leq \check{l}_K(\check{\theta}) - E[l_{K0}(\check{\theta})] + o_p(1) \leq \sup_{\theta\in\Theta}|\check{l}_K(\theta) - E[l_{K0}(\theta)]| + o_p(1) = o_p(1). \tag{12}$$

Now take θ such that $|\theta-\theta_0| \geq \varepsilon$ for any fixed $\varepsilon > 0$. By (11) there must exist some $\gamma_\varepsilon > 0$ such that

$$E[l_{K0}(\check{\theta})] + \gamma_\varepsilon < E[l_{K0}(\theta_0)].$$

It follows that

$$P(|\check{\theta}-\theta_0| \geq \varepsilon) \leq P(E[l_{K0}(\check{\theta})] + \gamma_\varepsilon < E[l_{K0}(\theta_0)]).$$

Equation (12) implies that

$$P(E[l_{K0}(\check{\theta})] + \gamma_\varepsilon < E[l_{K0}(\theta_0)]) \to 0 \text{ as } K \to \infty.$$

Therefore

$$P(|\check{\theta}-\theta_0| \geq \varepsilon) \to 0 \text{ as } K \to \infty$$

which proves the consistency of $\check{\theta}$.

Proof of Theorem 3. Take a first order Taylor series expansion of $\hat{U}_K(\hat{\theta})$ around and θ_0:

$$\hat{U}_K(\hat{\theta}) = \hat{U}_K(\theta_0) + (\hat{\theta}-\theta_0)\left.\frac{\partial \hat{U}_K}{\partial\theta}\right|_{\theta=\theta^*} \tag{13}$$

where θ^* is between $\hat{\theta}$ and θ_0. It must be the case that $\hat{U}_K(\hat{\theta}) = 0$ since $\hat{\theta}$ was taken to be the maximum of $L(\theta;\check{\boldsymbol{\beta}},\check{\Lambda})$. Therefore

$$\sqrt{K}(\hat{\theta}-\theta_0) = \frac{\sqrt{K}\hat{U}_K(\theta_0)}{-\left.\frac{\partial\hat{U}_K}{\partial\theta}\right|_{\theta=\theta^*}}. \tag{14}$$

We already showed that $\hat{\theta}$ consistently estimates θ_0, so the law of large numbers implies that

$$\left.\frac{\partial \hat{U}_K}{\partial \theta}\right|_{\theta=\theta^*} \xrightarrow{P} W(\theta_0) = \lim_{K\to\infty} \left.\frac{\partial U_K}{\partial \theta}\right|_{\theta=\theta_0} \quad \text{(Fisher information)}.$$

We will show that the score equation $\hat{U}_K(\theta_0)$ in the numerator of (14) follows a normal distribution. Hereto we need a Taylor series expansion of $\hat{U}_K(\theta_0)$ around $\boldsymbol{\beta}_0$ and Λ_0. Because Λ_0 is an unspecified function, we will use the Hadamard derivative of $U_K(\theta_0)$ w.r.t. Λ at $\Gamma - \Lambda \in BV[0,\tau]$.

$$\frac{d}{dt}\left[K^{-1}\frac{\partial \log L(\theta;\boldsymbol{\beta},\Lambda + t(\Gamma-\Lambda))}{\partial\theta}\right]\Bigg|_{t=0} = \int_0^\tau \xi_K(\theta;\Lambda)(u)d(\Gamma-\Lambda)(u)$$

where

$$\xi_K(\theta;\Lambda)(u) = K^{-1}\sum_{i=1}^{K}\sum_{j=1}^{n_i} D_{ij}^U Y_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}(u)]$$

and

$$\begin{aligned} D_{ij}^U = \Bigg\{ & \delta_{ij}\frac{\varphi_\theta''\left(\varphi_\theta^{-1}(H_{ij})\right)}{\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)}\frac{\partial}{\partial\theta}\left[\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)\right]^{-1} \\ & + \delta_{ij}\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)\frac{\partial}{\partial\theta}\left[-\frac{\varphi_\theta''\left(\varphi_\theta^{-1}(H_{ij})\right)}{\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)^3}\right] \\ & - \frac{\varphi_\theta^{(d_i+1)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)}{\left[\varphi_\theta^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)\right]^2}\frac{1}{\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)}\frac{\partial}{\partial\theta}\left[\varphi_\theta^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)\right] \\ & + \frac{1}{\varphi_\theta^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)}\frac{\partial}{\partial\theta}\left[\frac{\varphi_\theta^{(d_i+1)}\left(\sum_{j=1}^{n_i}\varphi_\theta^{-1}(H_{ij})\right)}{\varphi_\theta'\left(\varphi_\theta^{-1}(H_{ij})\right)}\right]\Bigg\}(-H_{ij}). \end{aligned}$$

The derivative of $U_K(\theta)$ w.r.t. $\boldsymbol{\beta}$ is given by

$$\xi_K(\theta;\boldsymbol{\beta}) = K^{-1}\sum_{i=1}^{K}\sum_{j=1}^{n_i} D_{ij}^U\int_0^\tau Y_{ij}(u)\boldsymbol{Z}_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}(u)]d\Lambda(u).$$

We require $||\xi_K(\theta;\Lambda)||_\infty$ and $||\xi_K(\theta;\boldsymbol{\beta})||$ to be bounded. By condition C5, the terms unique to $\xi_K(\theta;\Lambda)$ and $\xi_K(\theta;\boldsymbol{\beta})$, i.e.

$$||Y_{ij}\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij}]||_\infty \quad \text{and} \quad \left|\left|\int_0^\tau Y_{ij}(u)\boldsymbol{Z}_{ij}(u)\exp[\boldsymbol{\beta}'\boldsymbol{Z}_{ij(u)}d\Lambda(u)]\right|\right|$$

are bounded. The common term $||D_{ij}^U||_\infty$ is also bounded.

A Taylor series expansion of $\hat{U}_K(\theta_0)$ around $\boldsymbol{\beta}_0$ and Λ_0 gives

$$\hat{U}_K(\theta_0) = U_{K0}(\theta_0) + \xi_K(\theta_0;\boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \int_0^\tau \xi_K(\theta_0;\Lambda_0)(t)d[\check{\Lambda}(t) - \Lambda_0(t)] + G_K,$$

where G_K is the remainder term for the Taylor series. Since $\check{\Lambda}$ is $\sqrt{K}$-consistent it can be shown that $G_K = o_p(K^{-1/2})$.

Define the pointwise limit of $\xi_K(\theta,\Lambda)(t)$ as $\xi(\theta,\Lambda)(t)$ and denote $\xi(\theta;\boldsymbol{\beta}) = E[\xi_K(\theta;\boldsymbol{\beta})]$. Since $||\xi_K(\theta;\Lambda)||_\infty$ and $||\xi_K(\theta;\boldsymbol{\beta})||$ are bounded, $||\xi(\theta;\Lambda)||_\infty$ and $||\xi(\theta;\boldsymbol{\beta})||$ are too. Therefore

$$\sqrt{K}\hat{U}_K(\theta_0) = \sqrt{K}\left(U_{K0}(\theta_0) + \xi(\theta_0;\boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \int_0^\tau \xi(\theta_0;\Lambda_0)(t)d[\check{\Lambda}(t) - \Lambda_0(t)]\right) + o_p(1). \tag{15}$$

By Spiekerman and Lin (1998)

$$\sqrt{K}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \boldsymbol{A}^{-1}\sum_{i=1}^{K}\boldsymbol{w}_{i\cdot}$$

where $\boldsymbol{w}_{i.}$ is the i^{th} component of the score function for $\boldsymbol{\beta}$ under the independence working assumption, evaluated at $\boldsymbol{\beta}_0$:

$$\boldsymbol{w}_{i.} = \sum_{j=1}^{n_i} \int_0^\tau \{\boldsymbol{Z}_{ij}(u) - E(\boldsymbol{\beta}_0, u)\} dM_{ij}(u)$$

with

$$M_{ij}(t) = \delta_{ij} Y_{ij}(t) - \int_0^t Y_{ij}(u) \exp\left[\boldsymbol{\beta}_0' \boldsymbol{Z}_{ij}(u)\right] d\Lambda_0(u).$$

They also showed that

$$\sqrt{K}(\check{\Lambda}_0(t, \check{\boldsymbol{\beta}}) - \Lambda_0(t)) \to \mathcal{W}(t) = K^{-1/2} \sum_{i=1}^{K} \Psi_i(t)$$

where $\mathcal{W}(t)$ is a zero-mean Gaussian process with variance function

$$E\left[\Psi_1(t)^2\right]$$

with

$$\Psi_i(t) = \int_0^t \frac{dM_{i.}(u)}{s^{(0)}(\boldsymbol{\beta}_0, u)} + \boldsymbol{h}^T(t) \boldsymbol{A}^{-1} \boldsymbol{w}_{i.}$$

and

$$\boldsymbol{h}(t) = -\int_0^t \boldsymbol{e}(\boldsymbol{\beta}_0, u) d\Lambda_0(u).$$

That's why

$$\begin{aligned}
&\sqrt{K}\left(U_{K0}(\theta_0) + \xi(\theta_0; \boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \int_0^\tau \xi(\theta_0; \Lambda_0)(t) d[\check{\Lambda}(t) - \Lambda_0(t)]\right) \\
&= \sqrt{K}\left(K^{-1} \sum_{i=1}^{K} \phi_i(\theta_0) + \xi(\theta_0; \boldsymbol{\beta}_0) K^{-1} \boldsymbol{A}^{-1} \sum_{i=1}^{K} \boldsymbol{w}_{i.} + \int_0^\tau \xi(\theta_0; \Lambda_0)(t) d\left[K^{-1} \sum_{i=1}^{K} \Psi_i(t)\right]\right) \\
&= K^{-1/2} \sum_{i=1}^{K} \left(\phi_i(\theta_0) + \xi(\theta_0; \boldsymbol{\beta}_0) \boldsymbol{A}^{-1} \boldsymbol{w}_{i.} + \int_0^\tau \xi(\theta_0; \Lambda_0)(t) d\Psi_i(t)\right) \\
&= K^{-1/2} \sum_{i=1}^{K} \Xi_i.
\end{aligned}$$

The central limit theorem implies that $\sqrt{K}\hat{U}_K(\theta_0)$ converges to a normally distributed random variable with mean zero and variance equal to the variance of Ξ_1.

Thus we have

$$\sqrt{K}(\hat{\theta} - \theta_0) = \frac{\sqrt{K}\hat{U}_K(\theta_0)}{-\left.\frac{\partial \hat{U}_K}{\partial \theta}\right|_{\theta=\theta^*}} \tag{16}$$

where

$$\sqrt{K}\hat{U}_K(\theta_0) \xrightarrow{D} N(0, \text{Var}(\Xi_1))$$

and

$$\left.\frac{\partial \hat{U}_K}{\partial \theta}\right|_{\theta=\theta^*} \xrightarrow{P} W(\theta_0).$$

By Slutsky's theorem, $\sqrt{K}(\hat{\theta} - \theta_0)$ converges to a normal distribution with mean zero and variance equal to

$$\frac{\text{Var}(\Xi_1)}{W(\theta_0)^2}.$$

The variance of Ξ_1 (note that $\text{Var}(\Xi_1) = E[\Xi_1^2]$) can be estimated by $K^{-1} \sum_{i=1}^{K} \hat{\Xi}_i^2$ where $\hat{\Xi}_i$ is obtained from Ξ_i replacing parameter values by their estimators.

$W(\theta_0)$ can be estimated by the (minus) derivative of the pseudo score function $\hat{U}_K(\theta)$, evaluated in $\hat{\theta}$.