Investigating the correlation structure of quadrivariate udder infection times through hierarchical Archimedean copulas

DOI: 10.1007/s10985-017-9411-6
Handle: http://hdl.handle.net/1942/25904

Noname manuscript No.
(will be inserted by the editor)

Investigating the correlation structure of quadrivariate udder infection times through hierarchical Archimedean copulas

the date of receipt and acceptance should be inserted later

Abstract The correlation structure imposed on multivariate time to event data is often of a simple nature, such as in the shared frailty model where pairwise correlations between event times in a cluster are all the same. In modeling the infection times of the four udder quarters clustered within the cow, more complex correlation structures are possibly required, and if so, such more complex correlation structures give more insight in the infection process. In this article, we will choose a marginal approach to study more complex correlation structures, therefore leaving the modeling of marginal distributions unaffected by the association parameters. The dependency of failure times will be induced through copula functions. The methods are shown for (mixtures of) the Clayton copula, but can be generalized to mixtures of Archimedean copulas for which the nesting conditions are met (McNeil, 2008; Hofert, 2011).

Keywords Quadrivariate event times · Archimedean copula · Mastitis · correlation structures

1 Introduction

Time to event data are often clustered and different techniques have been developed to cope with the clustering in the data. Most commonly used approaches only accommodate a simple association structure between the event times in a cluster. For instance, the underlying assumption of the shared frailty model is that the correlation between any two event times is the same (Duchateau and Janssen, 2008). The correlated frailty model allows more complex structures, but has mostly been used to model bivariate survival data (Wienke, 2011), and the extension of the correlated frailty model based on the gamma density function imposes quite a few restrictions on the correlation structure. An alternative modeling technique is based on copula functions.

Address(es) of author(s) should be given

Copulas have also been mostly used for bivariate survival data, and in the case clusters were larger, the development was most often also restricted to simple association structures. The data set studied here warrants the development of more complex association structures. We investigate the appropriateness of different association structures for the quadrivariate udder quarter infection times clustered in the cow. It was shown in previous analyses using frailty models with a simple association structure that strong correlation exists between the infection times within an udder (Goethals et al, 2009; Ampe et al, 2012). The udder quarters, however, can be ranked in space, and special correlation structures can therefore be proposed. For instance, it is biologically plausible that infection times of left and right udder quarters in front are more correlated than the right front and rear udder quarters. The method proposed here provides the tools to test such biologically plausible hypotheses. The findings have large impact on the prevention of infections in udder quarters, as large correlations could signify that bacteria are spread from one udder quarter to the next, which could be prevented with proper hygienic measures. In this article, we will use hierarchical Archimedean copula models as it allows us to impose the correlation structures with biological relevance to the quadrivariate udder quarter infection times. One challenge in using these types of models in multivariate survival data typically is that one needs to calculate all possible first and higher order partial derivatives of the joint survival function, which can get complicated if you allow for hierarchical structures.

In Section 2, we discuss the infection time data in cow udders and the general construction of the likelihood function. In Section 3, we introduce models with different correlation structures and also discuss the choice of the baseline hazard function and the one- and two-stage approach to model fitting. In Section 4, we describe the results of the different models and compare them with each other. There is a higher level of association within the two front and the two rear udder parts, than between pairs where one part is located front and one is located rear. There is no difference in association between infection times in multiparous and primiparous cows. Size and power calculations are performed in Section 5.

2 Time to infection data and the general likelihood function for a cow udder

We investigate the correlation structure between the times to infection of the four udder quarters nested in a cow. In total, 1196 cows have been followed up during the lactation period, which is roughly 300-350 days but different for every cow. We define time to infection, expressed in trimesters, as the midpoint between the sampling times of the last negative result and the first positive result. We model the time until infection with any bacteria, with the cow being the cluster and the quarter the experimental unit within the cluster. Observations are right-censored if no infection occurs before the end of the lactation period, or if the cow is lost to follow-up during the study. The censoring percentage is 61%. As a covariate, we consider the parity of the cow, since several studies have shown that prevalence as well as incidence of intramammary infections increase with parity (Weller et al, 1992). This covariate acts on the cow level. The parity is 0 for primiparous cows (cows that had one calving) and is 1 for multiparous cows (cows that had more than one calving). This data set is an example of a balanced design, since all clusters have four components. We order the four parts in each cluster such that each first component corresponds to the left-front udder quarter, each second component to the right-front udder quarter, each third component to the left-rear udder quarter and each fourth component to the right-rear part. Let K be the number of cows ($i = 1, \ldots, K$). In each udder, we denote the lifetime for the different parts by a positive random variable T_{ij}, $j = 1, \ldots, 4$. For each cow, we assume that there is an independent random censoring variable C_i such that under a right censoring scheme, the observed quantities are given by $X_{ij} = \min(T_{ij}, C_i), \delta_{ij} = I(T_{ij} \leq C_i), i = 1, \ldots, K, j = 1, \ldots, 4$. The risk of infection may also depend on a set of covariates $\boldsymbol{Z}_{ij}$. As denoted in the previous section, we will consider the parity as a covariate, writing $Z_i = 0$ for primiparous cows and $Z_i = 1$ for multiparous cows. We denote the (possibly unobserved) time until infection for the different udder quarters of cow i by $(t_{i1}, t_{i2}, t_{i3}, t_{i4})$. The likelihood contribution of a cluster of size 4 is one term out of 16 possibilities. If cow i has four infected udder parts, its contribution to the likelihood is the joint density function of the infection times $f(t_{i2}, t_{i2}, t_{i3}, t_{i4}|Z_i)$. If the cow has only one infected udder part, we need to take the derivative of the joint survival function with respect to that event time, and so on. A general expression for the full likelihood is given by (1). Denote $\mathbf{x}_i = (x_{i2}, x_{i2}, x_{i3}, x_{i4})$.

$$\begin{aligned}
&\prod_{i=1}^{K} (f(\mathbf{x}_i|Z_i))^{\delta_{i1}\delta_{i2}\delta_{i3}\delta_{i4}} \\
&\quad \times \left(-\frac{\partial S(\mathbf{x}_i|Z_i)}{\partial x_{i1}}\right)^{\delta_{i1}(1-\delta_{i2})(1-\delta_{i3})(1-\delta_{i4})} \cdots \\
&\quad \times \left(\frac{\partial^2 S(\mathbf{x}_i|Z_i)}{\partial x_{i1}\partial x_{i2}}\right)^{\delta_{i1}\delta_{i2}(1-\delta_{i3})(1-\delta_{i4})} \cdots \\
&\quad \times \left(-\frac{\partial^3 S(\mathbf{x}_i|Z_i)}{\partial x_{i1}\partial x_{i2}\partial x_{i3}}\right)^{\delta_{i1}\delta_{i2}\delta_{i3}(1-\delta_{i4})} \cdots \\
&\quad \times S(\mathbf{x}_i|Z_i))^{(1-\delta_{i1})(1-\delta_{i2})(1-\delta_{i3})(1-\delta_{i4})}
\end{aligned} \tag{1}$$

3 Different models for the association structure

In this section, we progress from models with simple correlation structure to models with more complex correlation structure. As the models are nested, likelihood ratio testing can be used to test whether such more complex correlation structures are required. Association structures are expressed in terms of the copula function, i.e., correlation is introduced through the copula function that links the marginal survival functions into the joint survival function.

We model covariate effects under the assumption of proportional hazards. The hazard of infection at time t for udder quarter j with covariate information Z is $h_j(t|Z) = h_{0j}(t)\exp(\beta Z)$. In this model, $h_{0j}(t)$ is the baseline hazard function that describes the hazard for udder quarter j of primiparous cows and $\exp(\beta)$ is the proportionate increase or reduction in hazard for multiparous cows. The marginal survival function is $S_j(t|Z) = S_{0j}(t)^{\exp(\beta Z)}$ where $S_{0j}(t) = \exp(-H_{0j}(t))$ is the baseline survival function of udder quarter j and H_{0j} is the cumulative hazard function. Different assumptions about the baseline survival function (or hazard function) lead to different kinds of proportional hazards models. One can assume a parametric form of the baseline survival, but Cox (1972) observed that inference about the covariate effects is also possible when there is no assumption at all on the baseline survival (or hazard) function. In all models that will follow, both a parametric and semiparametric approach will be considered. In the parametric approach, the baseline hazard is assumed to be Weibull as this was shown to be an adequate choice by Goethals et al (2009). The general likelihood function (1) can then be maximized in one stage, i.e., maximizing the likelihood jointly for the parameter(s) of the marginal survival functions and the parameter in the copula function, or in two stages, first estimating the parameter(s) of the marginal survival functions, plugging those in (1) and then maximizing only for the parameter in the copula function. In the semiparametric approach, the baseline hazard is unspecified and only a two-stage approach is feasible: partial likelihood maximization is used to estimate the marginal survival functions, and only the association parameter remains in (1), for which it needs to be maximized. The two-stage approach is straightforward as the first stage, estimating the marginal survival functions, is based on basic survival models without clustering, and only one parameter remains in (1) for which it needs to be maximized.

3.1 No clustering (model 0)

In the case of independence between the udder quarter infection times in a cow, the joint survival function is given by

$$S(t_1, t_2, t_3, t_4|Z) = S_1(t_1|Z)S_2(t_2|Z)S_3(t_3|Z)S_4(t_4|Z)$$

where $S(t_1, t_2, t_3, t_4|Z)$ is the joint survival function and $S_1(t_1|Z), \ldots, S_4(t_4|Z)$ are the marginal survival functions for the left front, right front, left rear and right rear udder quarters respectively.

3.2 One level of clustering (model 1)

If we assume that the association between each two different udder quarters is the same, we model the joint survival function by a four-dimensional Archimedean copula function with generator φ. The joint survival function is represented as

$$S(t_1, t_2, t_3, t_4|Z) = C_{\theta_0}(S_1(t_1|Z), S_2(t_2|Z), S_3(t_3|Z), S_4(t_4|Z)),$$

or equivalently,

$$S(t_1, t_2, t_3, t_4|Z) = \varphi\left[\varphi^{-1}(S_1(t_1|Z)) + \varphi^{-1}(S_2(t_2|Z)) + \varphi^{-1}(S_3(t_3|Z)) + \varphi^{-1}(S_4(t_4|Z))\right]$$

where $\varphi : [0, \infty[\rightarrow [0, 1]$ is a continuous, strictly decreasing function which is completely monotonic and has $\varphi(0) = 1$ and $\varphi(\infty) = 0$ (Nelsen, 2006). The generator φ depends on the association parameter θ_0. As an example, the association structure is induced here through a Clayton copula with generator $\varphi(t) = (1 + \theta_0 t)^{-1/\theta_0}$ with $\theta_0 > 0$. Infection times are independent when θ_0 approaches zero. For the case of a parametric baseline hazard and using the one-stage procedure, the contributions to the likelihood expression (1) are derived in the Appendix. Prenen et al (2017) show that maximizing the likelihood expression is equivalent to solving $\dfrac{d\log L}{d\boldsymbol{\eta}} = \mathbf{0}$ with

$$L = \prod_{i=1}^{K}\left(\prod_{j=1}^{n_i}\left[\frac{f_j(x_{ij}|Z_i)}{\varphi'(\varphi^{-1}(S_j(x_{ij}|Z_i)))}\right]^{\delta_{ij}}\right)\varphi^{(d_i)}\left(\sum_{j=1}^{n_i}\varphi^{-1}(S_j(x_{ij}|Z_i))\right). \quad (2)$$

where $n_i = 4$, $f_j(x_{ij}|Z_i) = -\dfrac{dS_j(x_{ij}|Z_i)}{dx_{ij}}$, $d_i = \sum_{j=1}^{n_i}\delta_{ij}$ and $\varphi^{(d_i)}$ is the derivative of order d_i of the generator φ. The parameter vector $\boldsymbol{\eta}$ contains the baseline parameters from the four margins, the parity effect β and the association parameter θ_0. Optimization can be done through standard numerical algorithms, e.g. using the R functions `optim` or `nlm`. Note that formula (2) is not restricted to the Clayton copula, but that it is a generic formula for the likelihood expression for any Archimedean copula with generator φ.

In the model above, it was not only assumed that all cows in the population can be described by the same correlation structure, but also that the correlations themselves are the same. As primiparous and multiparous cows react quite differently with respect to udder quarter infections, it is worthwhile to test whether primiparous and multiparous cows share the same values for the correlations within the same correlation structure. To test whether the association between infection times depends on the parity of the cow, we use the following copula function $\varphi(t) = (1 + \theta_p t)^{-1/\theta_p}$ for primiparous cows and $\varphi(t) = (1 + \theta_m t)^{-1/\theta_m}$ for multiparous cows and test the hypothesis

$$H_0 : \theta_m = \theta_p \quad \text{versus} \quad H_1 : \theta_m \neq \theta_p$$

which can then be tested through the likelihood ratio test.

3.3 Multilevel clustering: parent copula with two identical child copulas (Model 2)

We assume that the front udder quarters have the same association as the rear udder quarters. Another type of association occurs between the front and rear udder quarters. This type of association structure is captured by a partially nested Archimedean copula function where the parent copula C_{θ_0} has two identical child copulas C_{θ_1} and C_{θ_1}:

$$C_{\theta_0}\left[C_{\theta_1}(S_1(t_1|Z), S_2(t_2|Z)), C_{\theta_1}(S_3(t_3|Z), S_4(t_4|Z))\right],$$

or equivalently,

$$\begin{aligned} S(t_1,t_2,t_3,t_4|Z) = \varphi_0 &\left[\varphi_0^{-1} \circ \varphi_1 \left\{\varphi_1^{-1}(S_1(t_1|Z)) + \varphi_1^{-1}(S_2(t_2|Z))\right\}\right. \\ &\left. + \varphi_0^{-1} \circ \varphi_1 \left\{\varphi_1^{-1}(S_3(t_3|Z)) + \varphi_1^{-1}(S_4(t_4|Z))\right\}\right]. \end{aligned}$$

The generator φ_0 describes the association between front and rear udder quarters, while φ_1 describes the association within front udder quarters and within rear udder quarters. According to McNeil (2008), for a general nested Archimedean structure to be a proper copula, it is sufficient that all appearing nodes of the form $\varphi_k^{-1} \circ \varphi_l$ have completely monotone derivatives. The sufficient nesting condition is often easily verified if all generators appearing in the nested structure come from the same parametric family. For the Archimedean families of Ali-Mikhail-Haq, Clayton, Frank, Gumbel and Joe, two generators φ_k and φ_l of the same family with corresponding parameters θ_k and θ_l fulfill the sufficient nesting condition if $\theta_k \leq \theta_l$ (Hofert, 2011). Choosing parent copula C_{θ_0} and child copulas C_{θ_1} to be Clayton copulas with generators $\varphi_0(t) = (1+\theta_0 t)^{-1/\theta_0}$ and $\varphi_1(t) = (1+\theta_1 t)^{-1/\theta_1}$ leads to the joint survival function

$$\begin{aligned} S(t_1,t_2,t_3,t_4|Z) = \Big[&-1 + \left(-1 + S_1(t_1|Z)^{-\theta_1} + S_2(t_2|Z)^{-\theta_1}\right)^{\theta_0/\theta_1} \\ &+ \left(-1 + S_3(t_3|Z)^{-\theta_1} + S_4(t_4|Z)^{-\theta_1}\right)^{\theta_0/\theta_1}\Big]^{-1/\theta_0}. \end{aligned}$$

The nesting condition for this setting is $\theta_0 \leq \theta_1$. This means that there must be a stronger association of infection times within front and within rear udder parts, than there is between front and rear udder parts. The contributions to the likelihood expression (1) for the one-stage parametric approach are given in the Appendix.

3.4 Multilevel clustering: parent copula with two different child copulas (model 3)

We will now assume that the association within the front udder quarters is different from the association within the rear udder quarters. A third type of association occurs between front and rear udder quarters. The hierarchical Archimedean copula function that represents this situation is

$$C_{\theta_0}\left[C_{\theta_1}(S_1(t_1|Z), S_2(t_2|Z)), C_{\theta_2}(S_3(t_3|Z), S_4(t_4|Z))\right],$$

or equivalently,

$$\begin{aligned} S(t_1, t_2, t_3, t_4|Z) = \varphi_0 \left[\varphi_0^{-1} \circ \varphi_1 \left\{\varphi_1^{-1}(S_1(t_1|Z)) + \varphi_1^{-1}(S_2(t_2|Z))\right\}\right. \\ \left. + \varphi_0^{-1} \circ \varphi_2 \left\{\varphi_2^{-1}(S_3(t_3|Z)) + \varphi_2^{-1}(S_4(t_4|Z))\right\}\right]. \end{aligned}$$

The generator φ_0 describes the association between the front and rear udder quarters, while generators φ_1 and φ_2 describe the association within the front udder quarters and within the rear udder quarters, respectively. We will choose φ_0, φ_1 and φ_2 to be generators of Clayton copulas with association parameters θ_0, θ_1 and θ_2. In that case, the joint survival function is given by

$$\begin{aligned} S(t_1, t_2, t_3, t_4|Z) = \Big[-1 + \left(-1 + S_1(t_1|Z)^{-\theta_1} + S_2(t_2|Z)^{-\theta_1}\right)^{\theta_0/\theta_1} \\ + \left(-1 + S_3(t_3|Z)^{-\theta_2} + S_4(t_4|Z)^{-\theta_2}\right)^{\theta_0/\theta_2}\Big]^{-1/\theta_0} \end{aligned}$$

The nesting conditions are $\theta_0 \leq \theta_1$ and $\theta_0 \leq \theta_2$. The contributions to the likelihood function are given in the Appendix. Note that when one rather fits Models 2 and 3 with other nested Archimedean copulas, the expressions in the Appendix no longer hold, i.e., all likelihood contributions need to be recalculated.

4 Results

4.1 The marginal survival functions

When assuming a parametric form of the marginal survival functions, the Weibull distribution is a popular choice. Under the Weibull assumption, the marginal survival functions are

$$S_j(t|Z) = \exp(-\lambda_j t^{\rho_j} \exp(\beta Z)), \quad j = 1, \ldots, 4.$$

The parity Z is cow-specific and therefore we assume that the parity effect is the same in each of the four quarters ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$). In a model without clustering, standard survival methods yield the parameter estimates in Table 1 under Model 0. When no parametric baseline is assumed, a stratified Cox proportional hazards model (Cox, 1972) can be used, where the baseline hazard $h_{0j}(\cdot)$ is left unspecified:

$$h_j(t|Z) = h_{0j}(t) \exp(\beta Z)$$

and the Nelson-Aalen estimator (Nelson, 1972) is used for the survival function. In this (semi-parametric) model, the estimated parity effect is 0.407(0.051). Following either the parametric or semi-parametric approach, these parameter estimates are consistent and are used in the second stage of the (semi-)parametric two-stage estimation procedure of models 1, 2 and 3.

Table 1 Parameter estimates for the udder quarter data based on model 0 (no correlation), copula model 1 (all correlations the same), copula model 2 (different correlation between pairs of udder quarters in front or at the rear compared to correlation between any front and rear udder quarter) and copula model 3 (the same as copula model 2, but with different correlation in front or at the rear). The first columns are based on one-stage estimation, the last two columns refer to the two-stage estimation.

		1-stage parametric estimation						**2-stage estimation**	
								parametric	semiparametric
		Weibull baseline parameters				parity effect β	copula parameter(s)	copula parameter(s)	
		$j = 1$	$j = 2$	$j = 3$	$j = 4$				
Model 0	λ_j	0.111	0.117	0.095	0.103	0.418			
		(0.008)	(0.009)	(0.008)	(0.008)	(0.051)			
	ρ_j	1.321	1.270	1.325	1.280				
		(0.052)	(0.050)	(0.055)	(0.053)				
Model 1	λ_j	0.111	0.118	0.095	0.100	0.344	$\theta_0 = 3.184$	$\theta_0 = 2.938$	$\theta_0 = 3.227$
		(0.009)	(0.009)	(0.008)	(0.008)	(0.068)	(0.182)	(0.201)	(0.216)
	ρ_j	1.297	1.262	1.310	1.266				
		(0.0483)	(0.047)	(0.052)	(0.050)				
Model 2	λ_j	0.112	0.118	0.095	0.101	0.340	$\theta_0 = 3.050$	$\theta_0 = 2.825$	$\theta_0 = 3.110$
		(0.009)	(0.010)	(0.008)	(0.008)	(0.068)	(0.184)	(0.186)	(0.215)
	ρ_j	1.299	1.264	1.310	1.269(0.050)		$\theta_1 = 3.552$	$\theta_1 = 3.299$	$\theta_1 = 3.626$
		(0.048)	(0.047)	(0.052)	(0.050)		(0.221)	(0.232)	(0.263)
Model 3	λ_j	0.111	0.118	0.095	0.101	0.340	$\theta_0 = 3.048$	$\theta_0 = 2.821$	$\theta_0 = 3.107$
		(0.009)	(0.010)	(0.008)	(0.008)	(0.068)	(0.185)	(0.187)	(0.215)
	ρ_j	1.299	1.264	1.311	1.270		$\theta_1 = 3.589$	$\theta_1 = 3.363$	$\theta_1 = 3.674$
		(0.048)	(0.047)	(0.052)	(0.050)		(0.271)	(0.282)	(0.323)
							$\theta_2 = 3.513$	$\theta_2 = 3.231$	$\theta_2 = 3.575$
							(0.274)	(0.278)	(0.309)

4.2 Fitting a hierarchy of association structures

Models 0, 1, 2 and 3 are fitted and the parameter estimates are reported in Table 1. Corresponding standard errors are in brackets. The standard errors of the associaton parameters were determined using the grouped jackknife procedure (Lipsitz et al, 1994; Lipsitz and Parzen, 1996). As pointed out in the previous section, in two-stage estimation, the estimates of the baseline and the parity effect are equal to the estimates arising from the independence model. To investigate which association structure is most appropriate, we test the hypotheses

$$\begin{array}{l} H_0^A : \theta_0 = 0 \quad \text{versus } H_1^A : \theta_0 > 0 \quad \text{in model 1} \\ H_0^B : \theta_1 = \theta_0 \text{ versus } H_1^B : \theta_1 > \theta_0 \text{ in model 2} \\ H_0^C : \theta_2 = \theta_1 \text{ versus } H_1^C : \theta_2 \neq \theta_1 \text{ in model 3} \end{array}$$

In words, test A is used to detect the presence of clustering in the data. With test B we determine whether it is necessary to account for front and rear subclusters. Test C is used to detect a different level of association in the front and rear subclusters. For testing $H_0^A : \theta = 0$, we use a likelihood ratio test with a mixed chi-squared distribution, since the null hypothesis lies at the boundary of the parameter space (Duchateau et al, 2002). In Section 5.1, we take a closer look at this distribution. To test hypotheses H_0^B and H_0^C, the likelihood ratio statistic follows a $\chi^2(1)$ distribution. The likelihood ratio tests are performed for both one-stage and two-stage estimation procedures, yielding similar p-values within each test. The 3 resulting p-values for test A are all < 0.0001 which makes us conclude that there is in fact clustering of infection times. Tests B and C result in p-values of < 0.0002 and > 0.6, respectively, so the front and rear subclusters are detected, but there is no need to set up a model which includes a different level of association within each subcluster. The most appropriate model for the udder quarter infection times, is therefore model 2.

4.3 The association structure as a function of the parity covariate

As mentioned at the end of Section 3.2, it is worthwhile to test the hypothesis of homogeneity of association

$$H_0^Z : \theta_m = \theta_p \quad \text{versus} \quad H_1^Z : \theta_m \neq \theta_p$$

in model 1. None of the estimation procedures lead to a significant difference between θ_m and θ_p. For example, the one-stage estimates are $\hat{\theta}_m = 3.055$ and $\hat{\theta}_p = 3.515$ where the test for equality yields a p-value of 0.231. Consequently, there is no need to model the association structure as a function of the parity covariate.

4.4 Interpretation of the correlation structure

More insight can be gained in the correlation structure implied by the particular copula by investigating conditional survival probabilities. In this section, we compare the independence model, i.e., Model 0, with the most appropriate model, i.e., Model 2 and consider conditional survival probabilities of the different quarters given that an event has taken place at time x in the first quarter. For the independence model, the conditional survival probability at time $t = x + u$, with u the time since infection in quarter one, is given by $P(T_j > x + u | T_1 = x, T_2 > x, T_3 > x, T_4 > x, Z = z) = \exp(-\lambda_j((x+u)^{\rho_j} - x^{\rho_j})\exp(\beta z))$ for $j = 2, 3, 4$ and when assuming marginal Weibull baseline hazards. For model 2, different conditional survival probabilities occur for quarter 2 as compared to quarter 3 and 4. For quarter 2, the conditional survival probability at time $t = x + u$ corresponds to

$$
\begin{aligned}
&P(T_2 > x + u | T_1 = x, T_2 > x, T_3 > x, T_4 > x, Z = z) \\
&= \left[\frac{-1 + \left(-1 + S_1(x)^{-\theta_1} + S_2(x+u)^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3(x)^{-\theta_1} + S_4(x)^{-\theta_1}\right)^{\theta_0/\theta_1}}{-1 + \left(-1 + S_1(x)^{-\theta_1} + S_2(x)^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3(x)^{-\theta_1} + S_4(x)^{-\theta_1}\right)^{\theta_0/\theta_1}} \right]^{-1/\theta_0 - 1} \\
&\times \left[\frac{-1 + S_1(x)^{-\theta_1} + S_2(x+u)^{-\theta_1}}{-1 + S_1(x)^{-\theta_1} + S_2(x)^{-\theta_1}} \right]^{\frac{\theta_0}{\theta_1} - 1}
\end{aligned}
\tag{3}
$$

with $S_j(x) = \exp(-\lambda_j x^{\rho_j} \exp(\beta z))$. This expression simplifies for the third (or fourth) quarter to

$$
\begin{aligned}
&P(T_3 > x + u | T_1 = x, T_2 > x, T_3 > x, T_4 > x) \\
&= \left[\frac{-1 + \left(-1 + S_1(x)^{-\theta_1} + S_2(x)^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3(x+u)^{-\theta_1} + S_4(x)^{-\theta_1}\right)^{\theta_0/\theta_1}}{-1 + \left(-1 + S_1(x)^{-\theta_1} + S_2(x)^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3(x)^{-\theta_1} + S_4(x)^{-\theta_1}\right)^{\theta_0/\theta_1}} \right]^{-1/\theta_0 - 1}
\end{aligned}
\tag{4}
$$

The derivations of the conditional survival probabilities are given in the Appendix.

In Figure 1, we note that the conditional survival probabilities for both quarters 2 and 3 are much lower in copula model 2 as compared to the independence model. This means that when a cow has an infected udder quarter, it has a higher risk of getting an infection in a nearby udder quarter in copula model 2 than under the independence model. Furthermore, we see that the conditional survival probability for quarter 2 is always smaller than for quarter 3. This is due to the fact that the association between quarter 1 and 2 is larger than between quarter 1 and 3. These results are consistently shown for both primiparous as multiparous cows, and when the first infection happens one week into the lactation period or only after one trimester in the lactation period.

5 Size and power analysis

We simulate survival data that resemble the udder infection data. All subjects are sampled from the same marginal distribution, i.e., a Weibull distribution with parameters comparable to the estimated parameters of the udder infection data set: $\lambda = 0.11, \rho = 1.3, \beta = 0.4$. The censoring variable is Weibull distributed with $\rho_C = 1.3$ and $\lambda_C = 0.21$, yielding a censoring percentage around 61%. The aim is to assess the size and the power of the likelihood ratio tests when comparing the different association structures. We only investigate the performance of the two-stage parametric estimation procedure. In the first simulation setting, we simulate four-dimensional survival data sets with one level of clustering, and calculate the size and the power to detect departures from the independence model. In the second simulation setting, we simulate data from a two-level hierarchical copula model with two identical child copulas, and compute the size and power to detect the subclusters.

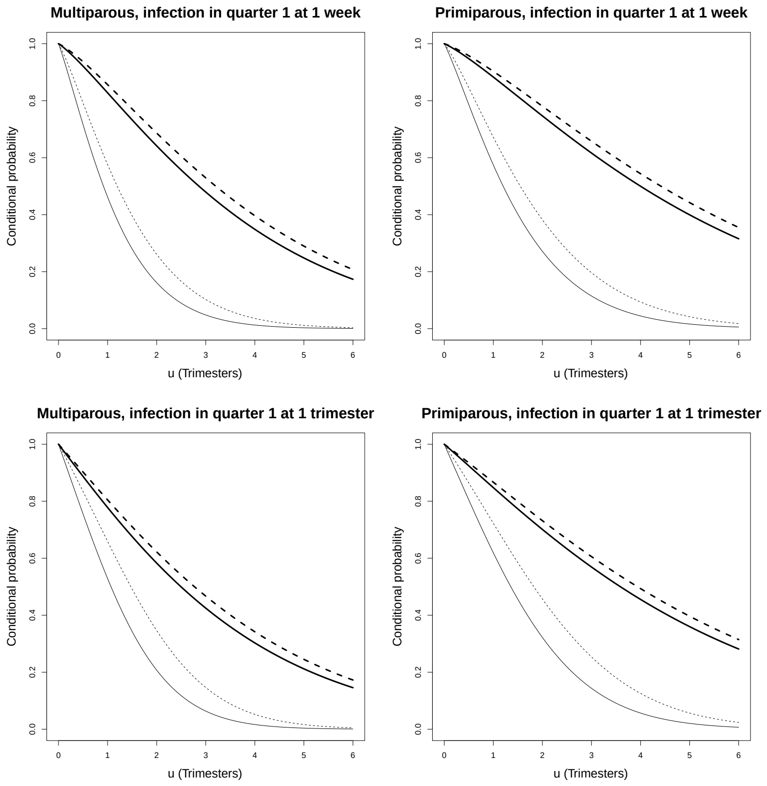

Fig. 1 Conditional survival probabilities in the 2nd (full line) and 3th (dashed) quarters of multiparous and heifer cows as a function of time since infection in the first quarter takes place at either 1 week or 1 trimester in the lactation period. The bold lines are the conditional probabilities under independence while the regular lines are obtained from model 2.

In the third simulation setting, data were simulated from a two-level hierarchical Archimedean copula model with two different child copulas, and the size and the power to detect the difference between the two subclusters were determined. In the last setting, we study the size and the power to detect a covariate effect on the association parameter in the model with one level of clustering.

5.1 Testing for independence versus one-level clustering

Let the true value of θ range from 0 to 0.5 by steps of size 0.05. We simulate 1000 data sets with 200 clusters of size 4 from a Clayton copula for each specific value of the association parameter θ. Our aim is to pick up deviations from independence. The power of the likelihood ratio test for independence is plotted versus the value of θ. At the boundary of the parameter space ($\theta = 0$), the likelihood ratio statistic follows a mixed chi-squared distribution

$$2 \log \frac{\text{likelihood alternative model}}{\text{likelihood null model}} \sim 0.5\chi^2(0) + 0.5\chi^2(1).$$

A value of $\theta = 0.20$, corresponding to a Kendall's tau of 0.09, is detected with a probability of 72.8%. In the model with one level of clustering, deviations from independence are hence quickly detected. From $\theta = 0.35$ onwards, the power level approaches 1. At $\theta = 0$, we approximately attain the size of the test by a value of 0.043.

5.2 Testing for one-level clustering versus two-level clustering with one parent and two identical child copulas

We let the true values of θ_0 and θ_1 range from 0.02 to 1 by steps of length 0.02, only considering those combinations of (θ_0, θ_1) for which the nesting condition $\theta_1 \geq \theta_0$ is met. We simulate 1000 data sets with 200 clusters of size 4 from a hierarchical Clayton copula with parent copula C_{θ_0} and 2 identical child copulas C_{θ_1} for each eligible pair (θ_0, θ_1) and calculate the probability to detect the subclusters. In order to make use of the $\chi^2(1)$ distribution, however, values of θ_0 and θ_1 close to the boundary, i.e., $\theta_0 = \theta_1 = 0$, should be excluded, as demonstrated in the discussion section. Therefore, first, the independence hypothesis is tested, and if not rejected, the simulated data set is discarded and not used in the future testing of $H_0^B : \theta_1 = \theta_0$. In simulation settings with θ_1 and θ_0 close to zero, a substantial number of the 1000 simulations might be discarded. If less than 20% of the 1000 data sets remain, a triangle symbol is used for plotting. As demonstrated in Section 5.1, this occurs only for very small values of the association parameter ($\theta_0 = \theta_1 \leq 0.06$). The size of the test is shown in the top panel of Figure 2; most values are below 0.06 and thus quite acceptable. In the bottom panel of Figure 2, the line $\theta_0 = \theta_1$ indicates the null model, i.e., no subclusters. To obtain a power of 80%, values must differ quite substantially, e.g., $(0.2, 0.75)$ or $(0.3, 0.9)$. In order to perform these tests, a sufficient number of clusters should be available, otherwise the size of the test will not be respected. The sample size used in our simulations, i.e., 200 clusters, is at the border and with fewer clusters the proposed tests should not be used.

5.3 Testing for two-level clustering with one parent and two identical child copulas versus two-level clustering with one parent and two different child copulas

We fix the value of θ_0 at 0.5 and let θ_1 and θ_2 range from 0.5 to 2.0 by steps of length 0.1. We simulate 1000 data sets with 200 clusters of size 4 from a hierarchical Clayton copula with parent copula C_{θ_0} and 2 child copulas C_{θ_1}

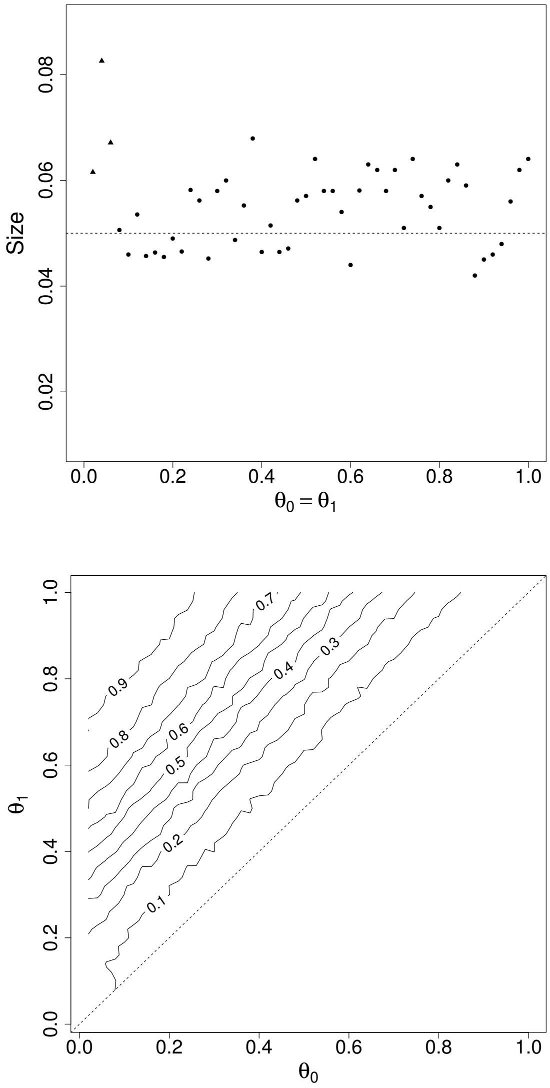

Fig. 2 Size (top) and power (bottom) of the likelihood ratio test for $H_0^B : \theta_0 = \theta_1$ versus $H_1^B : \theta_0 < \theta_1$.

and C_{θ_2}. For each combination of θ_1 and θ_2, we calculate the probability to detect the different levels of association in the subclusters. In Figure 3, the line $\theta_1 = \theta_2$ indicates the model with two identical child copulas. On this line, we achieve the size of the test.

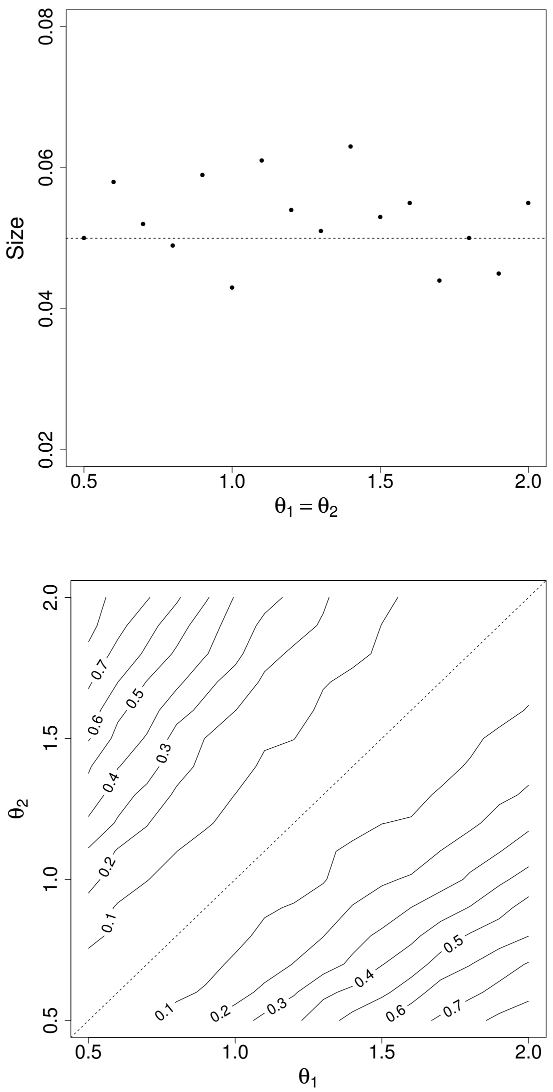

Fig. 3 Size (top) and power (bottom) of the likelihood ratio test for $H_0^C : \theta_1 = \theta_2$ versus $H_1^C : \theta_1 \neq \theta_2$.

5.4 Testing for differing association structures as a function of a covariate

We fix the true value of θ_p at 1.5 and let θ_m range from 0 to 3.5 by steps of length 0.1. We simulate 1000 data sets with 200 clusters of size 4 from a Clayton copula C_{θ_m} for a multiparous cow and from a Clayton copula C_{θ_p} for a primiparous cow. We determine how many times the difference between the

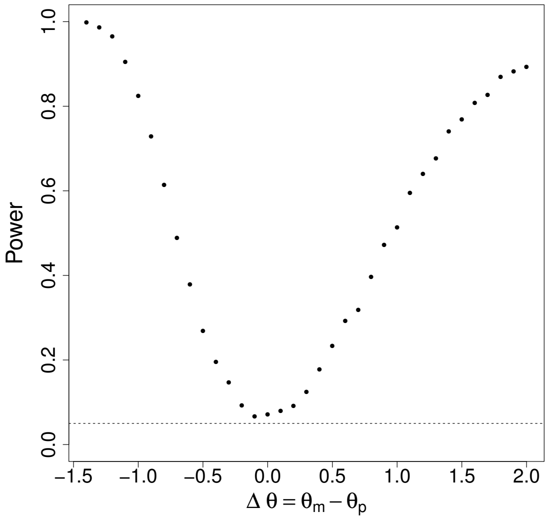

Fig. 4 The power of the likelihood ratio test for $H_0^Z : \theta_m = \theta_p$ versus $H_1^Z : \theta_m \neq \theta_p$.

association parameters θ_m and θ_p is picked up. In Figure 4, when $\Delta\theta = \theta_m - \theta_p$ approaches -1.5, i.e., when C_{θ_m} approaches the independence copula, the power increases quickly. On the right hand side of Figure 4, where $\Delta\theta$ is positive, the power to detect a covariate effect on the association parameter increases more gradually.

5.5 Robustness of the Clayton copula

To study the robustness of the Clayton copula model assumption against misspecification, we simulated 1000 data sets with 1000 clusters of size 4 from a Gumbel copula. To mimic the time to infection data, we generate survival times from a Weibull distribution with $\lambda = 0.11$ and $\rho = 1.3$. The covariate effect of a binary covariate (the parity) is set equal to $\beta = 0.34$. (These values roughly correspond to the values in Table 1.) The distribution of the censoring times is also Weibull with $\lambda_C = 0.20$ and $\rho_C = 1.3$, yielding a censoring percentage of 60%. Data were first generated for the association structure of Model 1, with only one level of association, and generator $\varphi_\theta(s) = \exp(-s^\theta)$ with $0 < \theta < 1$ where association gets stronger as θ approaches 0. We next fitted a Clayton copula to these data using the two-stage parametric procedure. In each data set, we estimate the Clayton copula parameter and calculate Kendall's tau using the formula $\tau = \frac{\theta}{\theta+2}$ as derived in Duchateau and Janssen (2008). If the Gumbel copula from which the data were generated, has an association parameter equal to 0.35, corresponding to a Kendall's tau of 0.65 ($\tau = 1 - \theta$), the mean of the 1000 estimated Kendall's taus using the Clayton copula, is 0.72 with a standard deviation of 0.015. With a Gumbel copula with an association parameter equal to 0.65, corresponding to a Kendall's tau of 0.35, the mean of the 1000 estimated Kendall's taus using the Clayton copula, is 0.43 with a standard deviation of 0.022. Next, we generated 1000 data sets with 1000 clusters of size 4 from a hierarchical Gumbel copula with two identical Gumbel child copulas (Model 2, nesting condition for Gumbel copulas is $\theta_0 \geq \theta_1$), and we fitted a hierarchical Clayton copula (Model 2) to these data. The theoretical values used for data generation in the Gumbel copula corresponded to $\theta_0 = 0.5$ ($\tau_0 = 0.5$) for the parent copula and $\theta_1 = 0.35$ ($\tau_1 = 0.65$) for the child Gumbel copulas. The means of the 1000 estimated Kendall's taus using the hierarchical Clayton copula equal $\tau_0 = 0.58$ (sd = 0.024) and $\tau_1 = 0.73$ (sd = 0.019). Finally, we generated 1000 data sets with 1000 clusters of size 4 from a hierarchical Gumbel copula with two different Gumbel child copulas (Model 3, nesting condition for Gumbel copulas). For the Gumbel copula we used $\theta_0 = 0.5$ ($\tau_0 = 0.5$) for the parent copula, and for the child Gumbel copulas $\theta_1 = 0.35$ ($\tau_1 = 0.65$) and $\theta_2 = 0.4$ ($\tau_2 = 0.6$). The means of the 1000 estimated Kendall's taus using the hierarchical Clayton copula equal $\tau_0 = 0.58$ (sd = 0.031), $\tau_1 = 0.73$ (sd = 0.045) and $\tau_2 = 0.68$ (sd = 0.049). We can conclude that the Kendall's tau estimate is biased upwards when data are generated from a Gumbel copula and analysed with a Clayton copula. Importantly, even in the presence of this upward bias, associations that are larger in the data generation based on Gumbel turn out to be larger as well in the analysis based on the Clayton copula.

6 Discussion

We compared different hierarchical Archimedean copula models for the association between infection times of the four udder parts in dairy cows. The most adequate model for the quadrivariate udder infection data is the nested copula model where the association between front and rear udder quarters is smaller than the association between two front, resp. rear, udder quarters. The within-front association is not significantly different from the within-rear association. According to the best fitting copula, i.e., Model 2, the association parameter between two quarters either on the rear or on the front side corresponds to 3.552, or a Kendall's tau equal to 0.64. As expected the association parameter between two quarters not on the same rear or front side is smaller and equal to 3.050 with a corresponding Kendall's tau equal to 0.60. Although these two association parameters differ significantly from each other, it is not important from a practical point of view as both are large. Note that the constraints on the parameter space were fulfilled for this particular data set, using standard R optimization algorithms. Whenever the constraints are not fulfilled for a particular data set, the R function will issue a warning message and propose a change to the nesting structure. The R function can be found in the **UdderQuarterInfectionData** package available from CRAN.

It is important to know for a dairy holder that noninfected udder quarters are highly at risk whenever one of the udder quarters of a cow is infected. A more straightforward approach to estimate correlations between udder quarter infection times is based on non-parametric estimation of Kendall's tau for censured time to event data. We used Hougaard's method (Hougaard, 2000, p.132) to estimate pairwise correlations between the different udder quarters. The average of all pairwise estimates equals 0.558, which is similar to the value of 0.617, i.e., $3.227/(2 + 3.227)$, obtained from the semiparametric version of Model 1. The average of Kendall's tau values for the pairs of udder quarters either in front or at the rear equals 0.591, whereas the average of Kendall's tau values for the pairs of udder quarters between any front and any rear quarter equals 0.558. Therefore, the Clayton copula and the nonparametric estimate of Kendall's tau give relatively the same values for Kendall's tau, but it seems that Kendall's tau values obtained from the Clayton copula are higher than those from the non-parametric estimation. The reason to choose all apparent copulas to be Clayton copulas is threefold. From a computational point of view, Clayton copulas are convenient to work with, since there exists a closed form expression for the derivatives of the copula generator φ. For a hierarchical Clayton copula to be well-defined, a simple nesting condition has to be met, i.e., $\theta_k \leq \theta_l$ for all appearing nodes of the form $\varphi_k^{-1} \circ \varphi_l$. Additionally, the Clayton copula has lower tail dependence. In a survival context, this translates to a stronger association later in time. It's therefore important to extend these copulas to other members of the Archimedean copula family to investigate when the correlation in time is strongest. If the data require the use of another (nested) Archimedean copula, model 1 can be fitted directly through the use of formula (2). For the nested models 2 and 3, one will need to recalculate all likelihood contributions, using the appropriate combination of generators φ_0, φ_1 and φ_2.

In Section 5.1, the power of testing for the presence of simple clustering, i.e., the same pairwise correlation between all udder quarter, in four-dimensional data was assessed using a likelihood ratio test with a mixed chi-squared distribution (Self and Liang, 1987). The null hypothesis of no association lies on the boundary of the parameter space, and thus the empirical cumulative distribution function of the likelihood ratio statistic, calculated for 1000 simulated datasets without clustering, agrees with the $0.5\chi^2(0) + 0.5\chi^2(1)$ distribution function, as depicted in the top panel of Figure 5. In Section 5.2, we assessed the power of testing for the presence of subclusters in four-dimensional data using a likelihood ratio test with a $\chi^2(1)$ distribution. The null hypothesis $H_0^B : \theta_0 = \theta_1$ lies on the boundary of the nesting condition $\theta_0 \leq \theta_1$, however, since the nesting condition only is sufficient and not necessary, the mixed chi-squared distribution does not apply unless $\theta_0 = \theta_1 = 0$. Before testing for multiple levels of clustering, it is therefore necessary to test first for the presence of simple clustering. Omitting this preliminary test can lead to test sizes much larger than the nominal significance level. We determined the size of the likelihood ratio test for $H_0^B : \theta_0 = \theta_1$ versus $H_1^B : \theta_0 < \theta_1$ for 1000 data sets with 200 clusters of size 4 that were simulated from a Clayton copula

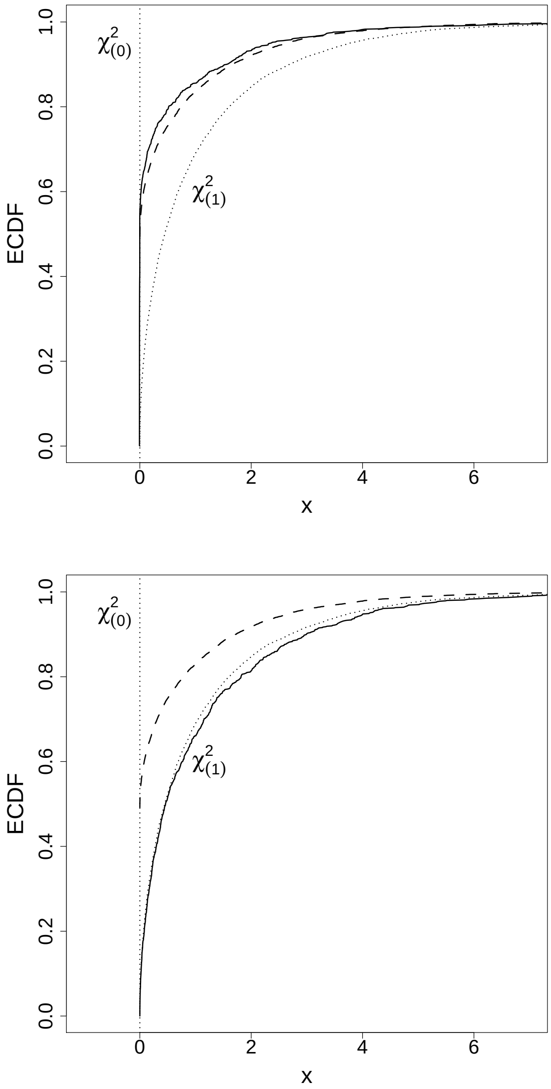

Fig. 5 Top: the empirical cumulative distribution function (ECDF) under $H_0^A : \theta = 0$ (solid line), ECDF of $\chi^2(0), \chi^2(1)$ (dotted lines) and $0.5\chi^2(0) + 0.5\chi^2(1)$ (dashed line). Bottom: ECDF under $H_0^B : \theta_0 = \theta_1 = 1.2$ (solid line), ECDF of $\chi^2(0), \chi^2(1)$ (dotted lines) and $0.5\chi^2(0) + 0.5\chi^2(1)$ (dashed line).

with association parameter ranging from 0.02 to 1. For small values of θ, the size deviates heavily from the desired 0.05 level. For $\theta = 0.02$, the according size was 0.14. This systematic overestimation of the test size occurred up to $\theta = 0.12$, as described in Figure 6. In the top panel of Figure 2 in Section 5.2,

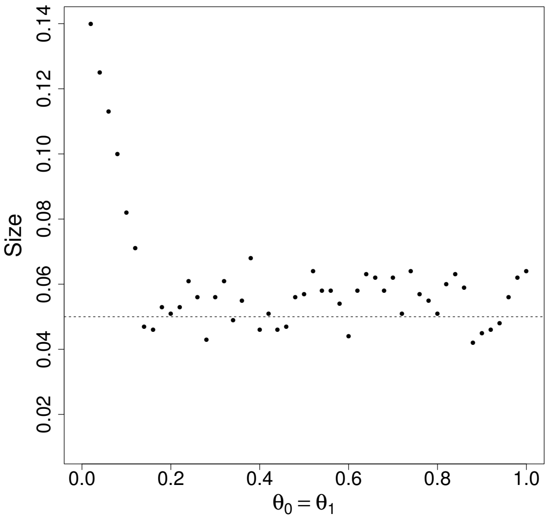

Fig. 6 The size of the likelihood ratio test for $H_0^B : \theta_0 = \theta_1$ versus $H_1^B : \theta_0 < \theta_1$, including also the data sets for which no simple clustering was detected.

we remedy this problem by only looking at those data sets in which the preliminary test detected the presence of simple clustering. The likelihood ratio test for $H_0^B : \theta_0 = \theta_1$ versus $H_1^B : \theta_0 < \theta_1$ was performed on 1000 data sets that were simulated from a unilevel Clayton copula model with association parameter equal to 1.2. The empirical cumulative distribution function of the likelihood ratio statistic nearly coincides with the $\chi^2(1)$ distribution, as can be seen in the bottom panel of Figure 5.

The Archimedean Clayton copula appears to be a useful tool to investigate the association structure in quadrivariate udder infection times, and an appealing interpretation of the association structure follows from considering the conditional survival probabilities. The studied models can be further extended in different ways. First, Laplace transforms other than the one linked to the Clayton copula, such as the inverse Gaussian or positive stable, could be considered. Alternative approaches such as vines (Barthel, 2015) or hierarchical Kendall copulas (Brechmann, 2014) could also be applied to these quadrivariate udder infection times. In future research, these alternative approaches will be compared to validate the association structure estimated in our model.

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 Economy, Science and Innovation.

References

Ampe B, Goethals K, Laevens H, Duchateau L (2012) Investigating clustering in interval-censored udder quarter infection times in dairy cows using a gamma frailty model. Preventive Veterinary Medicine 106(3-4):251–257

Barthel N (2015) Multivariate survival using vine-copulas. Master's thesis Technische Universitat Munchen

Brechmann EC (2014) Hierarchical kendall copulas: properties and inferences. The Canadian Journal of Statistics 42:78–108

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

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

Duchateau L, Janssen P, Lindsey P, Legrand C, Nguti R, Sylvester R (2002) The shared frailty model and the power for heterogeneity tests in multicenter trials. Computational Statistics and Data Analysis 40:603–620

Goethals K, Ampe B, Berkvens D, Laevens H, Janssen P, Duchateau L (2009) Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model. Journal of Agricultural Biological and Environmental Statistics 14(1):1–14

Hofert M (2011) Efficiently sampling nested archimedean copulas. Computational Statistics and Data Analysis 55:57–70

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

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

Lipsitz SR, Dear KB, Zhao L (1994) Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. Biometrics 50:842–846

McNeil AJ (2008) Sampling nested archimedean copulas. Journal of Statistical Computation and Simulation 6:567–581

Nelsen RB (2006) An Introduction to copulas. Springer

Nelson W (1972) Theory and applications of hazard plotting for censored failure data. Technometrics 14:945–965

Prenen L, Braekers R, Duchateau L (2017) Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. Journal of the Royal Statistical Society, Series B 79:483–505

Self S, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605–610

Weller J, Saran A, Zeliger Y (1992) Genetic and environmental relationships among somatic cell count, bacterial infection, and clinical mastitis. Journal of Dairy Science 75:2535–2540

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

A Calculation of likelihood contributions

The contributions to the likelihood (1) for the different association structures are given in this Appendix. For a sample of quadrivariate survival data

$$\{(x_{i1}, \delta_{i1}), (x_{i2}, \delta_{i2}), (x_{i3}, \delta_{i3}), (x_{i4}, \delta_{i4})\}, \quad i = 1, \ldots, K$$

the contribution L_i of quadruple i to the likelihood depends on the censoring status of the event times. In case of one-stage estimation, one needs to take derivatives with respect to the event time, whereas for two-stage estimation, it is sufficient to take derivatives at the level of the marginal survival functions. E.g., for one-stage estimation, the contributions to the likelihood are:

- For a cluster with no events: $L_i = S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)$
- For a cluster with one event: $L_i = \dfrac{\partial S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)}{\partial x_{ij}}$
- For a cluster with two events: $L_i = \dfrac{\partial^2 S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)}{\partial x_{ij} \partial x_{ik}}$
- For a cluster with three events: $L_i = \dfrac{\partial^3 S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)}{\partial x_{ij} \partial x_{ik} \partial x_{il}}$
- For a cluster with four events: $L_i = \dfrac{\partial^4 S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)}{\partial x_{i1} \partial x_{i2} \partial x_{i3} \partial x_{i4}}$

According to the association structure that is specific to Model 0, 1, 2 and 3, the joint survival function $S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)$ takes on a different form.

We denote the joint survival function $S = S(x_{i1}, x_{i2}, x_{i3}, x_{i4}|Z_i)$ and the marginal survival functions $S_j = S_j(x_{ij}|Z_i)$, $j = 1, 2, 3, 4$. Furthermore, $S'_j = \dfrac{dS_j}{dx_{ij}}$. In two-stage estimation, the S'_j can be omitted, since the marginal survival functions do not contain information on the association parameter.

A.1 Model 0

In the independence model, the joint survival function of the quadrivariate lifetimes is

$$S = S_1 S_2 S_3 S_4$$

with derivatives

$$\begin{aligned}
\frac{\partial S}{\partial x_{ij}} &= S'_j S_k S_l S_m \text{ for } \{j, k, l, m\} = \{1, 2, 3, 4\} \\
\frac{\partial^2 S}{\partial x_{ij} \partial x_{ik}} &= S'_j S'_k S_l S_m \\
\frac{\partial^3 S}{\partial x_{ij} \partial x_{ik} \partial x_{il}} &= S'_j S'_k S'_l S_m \\
\frac{\partial^4 S}{\partial x_{i1} \partial x_{i2} \partial x_{i3} \partial x_{i4}} &= S'_1 S'_2 S'_3 S'_4
\end{aligned}$$

A.2 Model 1

In the model with one level of clustering, the joint survival function of the quadrivariate lifetimes is

$$S = \varphi \left[\varphi^{-1}(S_1) + \varphi^{-1}(S_2) + \varphi^{-1}(S_3) + \varphi^{-1}(S_4)\right].$$

For the Clayton copula, $\varphi(t) = (1+\theta t)^{-1/\theta}$ and $\varphi^{-1}(t) = \frac{t^{-\theta}-1}{\theta}$, yielding

$$S = \left[S_1^{-\theta} + S_2^{-\theta} + S_3^{-\theta} + S_4^{-\theta} - 3\right]^{-1/\theta}. \tag{5}$$

Now put

$$\begin{aligned} A &= \left[S_1^{-\theta} + S_2^{-\theta} + S_3^{-\theta} + S_4^{-\theta} - 3\right] \\ C_j &= S_j^{-\theta-1} S_j' \quad j = 1,2,3,4, \end{aligned}$$

then

$$\begin{aligned} \frac{\partial S}{\partial x_{ij}} &= A^{-1/\theta-1} C_j \\ \frac{\partial^2 S}{\partial x_{ij} \partial x_{ik}} &= (1+\theta) A^{-1/\theta-2} C_j C_k \\ \frac{\partial^3 S}{\partial x_{ij} \partial x_{ik} \partial x_{il}} &= (1+\theta)(1+2\theta) A^{-1/\theta-3} C_j C_k C_l \\ \frac{\partial^4 S}{\partial x_{i1} \partial x_{i2} \partial x_{i3} \partial x_{i4}} &= (1+\theta)(1+2\theta)(1+3\theta) A^{-1/\theta-4} C_1 C_2 C_3 C_4 \end{aligned}$$

A.3 Model 2

In the model with a parent copula with two identical child copulas, the joint survival function of the quadrivariate lifetimes is

$$\begin{aligned} S(t_1,t_2,t_3,t_4) = \varphi_0 \Big[& \varphi_0^{-1} \circ \varphi_1 \left\{ \varphi_1^{-1}(S_1(t_1)) + \varphi_1^{-1}(S_2(t_2)) \right\} \\ & + \varphi_0^{-1} \circ \varphi_1 \left\{ \varphi_1^{-1}(S_3(t_3)) + \varphi_1^{-1}(S_4(t_4)) \right\} \Big]. \end{aligned}$$

For the Clayton copula, this becomes

$$S = \left[-1 + \left(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3^{-\theta_1} + S_4^{-\theta_1}\right)^{\theta_0/\theta_1} \right]^{-1/\theta_0}.$$

Now put

$$\begin{aligned} A &= \left[-1 + \left(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\right)^{\theta_0/\theta_1} + \left(-1 + S_3^{-\theta_1} + S_4^{-\theta_1}\right)^{\theta_0/\theta_1} \right] \\ B_{12} &= \left(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\right) \\ B_{34} &= \left(-1 + S_3^{-\theta_1} + S_4^{-\theta_1}\right) \\ C_j &= S_j^{-\theta_1-1} S_j' \quad j = 1,2,3,4 \end{aligned}$$

then

$$\begin{aligned} \frac{\partial S}{\partial x_{ij}} &= \begin{cases} A^{-1/\theta_0-1} B_{12}^{\theta_0/\theta_1-1} C_j & \text{if } j = 1,2 \\ A^{-1/\theta_0-1} B_{34}^{\theta_0/\theta_1-1} C_j & \text{if } j = 3,4 \end{cases} \\ \frac{\partial^2 S}{\partial x_{ij} \partial x_{ik}} &= \begin{cases} A^{-1/\theta_0-2} B_{ij}^{\theta_0/\theta_1-2} C_j C_k \left[(1+\theta_0) B_{jk}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) A\right] \\ \hfill \text{if } (j,k) = (1,2),(3,4) \\ (1+\theta_0) A^{-1/\theta_0-2} B_{12}^{\theta_0/\theta_1-1} B_{34}^{\theta_0/\theta_1-1} C_j C_k \text{ else} \end{cases} \\ \frac{\partial^3 S}{\partial x_{ij} \partial x_{ik} \partial x_{il}} &= \begin{cases} (1+\theta_0) A^{-1/\theta_0-3} B_{12}^{\theta_0/\theta_1-2} B_{34}^{\theta_0/\theta_1-1} C_j C_k C_l \left[(1+2\theta_0) B_{12}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) A\right] \\ \hfill \text{if } (j,k,l) = (1,2,3),(1,2,4) \\ (1+\theta_0) A^{-1/\theta_0-3} B_{12}^{\theta_0/\theta_1-1} B_{34}^{\theta_0/\theta_1-2} C_j C_k C_l \left[(1+2\theta_0) B_{34}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) A\right] \\ \hfill \text{if } (j,k,l) = (1,3,4),(2,3,4) \end{cases} \\ \frac{\partial^4 S}{\partial x_{i1} \partial x_{i2} \partial x_{i3} \partial x_{i4}} &= (1+\theta_0) A^{-1/\theta_0-4} B_{12}^{\theta_0/\theta_1-2} B_{34}^{\theta_0/\theta_1-2} C_1 C_2 C_3 C_4 \\ & \quad \cdot \Big[(1+2\theta_0)(1+3\theta_0) B_{12}^{\theta_0/\theta_1} B_{34}^{\theta_0/\theta_1} + (1+2\theta_0)(-\theta_0+\theta_1) A (B_{12}^{\theta_0/\theta_1} + B_{34}^{\theta_0/\theta_1}) \\ & \quad + (-\theta_0+\theta_1)^2 A^2\Big] \end{aligned}$$

A.4 Model 3

In the model with a parent copula with two different child copulas, the joint survival function of the quadrivariate lifetimes is, for the Clayton copula

$$S = \Big[-1 + \Big(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\Big)^{\theta_0/\theta_1} \\ + \Big(-1 + S_3^{-\theta_2} + S_4^{-\theta_2}\Big)^{\theta_0/\theta_2}\Big]^{-1/\theta_0}$$

Now put

$$\begin{aligned}
A &= \Big[-1 + \Big(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\Big)^{\theta_0/\theta_1} + \Big(-1 + S_3^{-\theta_2} + S_4^{-\theta_2}\Big)^{\theta_0/\theta_2}\Big] \\
B_{12} &= \Big(-1 + S_1^{-\theta_1} + S_2^{-\theta_1}\Big) \\
B_{34} &= \Big(-1 + S_3^{-\theta_2} + S_4^{-\theta_2}\Big) \\
C_{j1} &= S_j^{-\theta_1-1} S'_j \quad j = 1, 2 \\
C_{j2} &= S_j^{-\theta_2-1} S'_j \quad j = 3, 4
\end{aligned}$$

then

$$\frac{\partial S}{\partial x_{ij}} = \begin{cases} A^{-1/\theta_0-1} B_{12}^{\theta_0/\theta_1-1} C_{j1} & \text{if } j = 1, 2 \\ A^{-1/\theta_0-1} B_{34}^{\theta_0/\theta_2-1} C_{j2} & \text{if } j = 3, 4 \end{cases}$$

$$\frac{\partial^2 S}{\partial x_{ij}\partial x_{ik}} = \begin{cases} A^{-1/\theta_0-2} B_{12}^{\theta_0/\theta_1-2} C_{11} C_{21} \Big[(1+\theta_0) B_{12}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) A\Big] & \text{if } (j,k) = (1,2) \\ A^{-1/\theta_0-2} B_{34}^{\theta_0/\theta_2-2} C_{32} C_{42} \Big[(1+\theta_0) B_{34}^{\theta_0/\theta_2} + (-\theta_0+\theta_2) A\Big] & \text{if } (j,k) = (3,4) \\ (1+\theta_0) A^{-1/\theta_0-2} B_{12}^{\theta_0/\theta_1-1} B_{34}^{\theta_0/\theta_2-1} C_{j1} C_{k2} & \text{if } j \in \{1,2\} \text{ and } k \in \{3,4\} \end{cases}$$

$$\frac{\partial^3 S}{\partial x_{ij}\partial x_{ik}\partial x_{il}} = \begin{cases} (1+\theta_0) A^{-1/\theta_0-3} B_{12}^{\theta_0/\theta_1-2} B_{34}^{\theta_0/\theta_2-1} C_{j1} C_{k1} C_{l2} \Big[(1+2\theta_0) B_{12}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) A\Big] \\ \qquad \text{if } (j,k,l) = (1,2,3), (1,2,4) \\ (1+\theta_0) A^{-1/\theta_0-3} B_{12}^{\theta_0/\theta_1-1} B_{34}^{\theta_0/\theta_2-2} C_{j1} C_{k2} C_{l2} \Big[(1+2\theta_0) B_{34}^{\theta_0/\theta_2} + (-\theta_0+\theta_2) A\Big] \\ \qquad \text{if } (j,k,l) = (1,3,4), (2,3,4) \end{cases}$$

$$\begin{aligned}
\frac{\partial^4 S}{\partial x_{i1}\partial x_{i2}\partial x_{i3}\partial x_{i4}} &= (1+\theta_0) A^{-1/\theta_0-4} B_{12}^{\theta_0/\theta_1-2} B_{34}^{\theta_0/\theta_2-2} C_{11} C_{21} C_{32} C_{42} \\
&\cdot \Big[(1+2\theta_0)(1+3\theta_0) B_{12}^{\theta_0/\theta_1} B_{34}^{\theta_0/\theta_2} \\
&+ (1+2\theta_0) A((-\theta_0+\theta_2) B_{12}^{\theta_0/\theta_1} + (-\theta_0+\theta_1) B_{34}^{\theta_0/\theta_2}) \\
&+ (-\theta_0+\theta_1)(-\theta_0+\theta_2) A^2\Big]
\end{aligned}$$

B Derivation of the conditional probabilities

In the different models, we express the joint survival function as

$$S(t_1, t_2, t_3, t_4) = C(S_1(t_1), S_2(t_2), S_3(t_3), S_4(t_4))$$

in which the copula function C describes the association between the different udder quarters. Based on this expression, we look for the conditional probabilities of udder quarter 2 and 3 to have no infection for at least a time u after an infection was seen in udder quarter 1 at time x. Hereby, we also assume that none of the other udder quarter had an infection. The conditional probability for udder quarter 2 (T_2) is given by

$$P(T_2 > x+u | T_1 = x, T_2 > x, T_3 > x, T_4 > x) = \frac{P(T_1 = x, T_2 > x+u, T_3 > x, T_4 > x)}{P(T_1 = x, T_2 > x, T_3 > x, T_4 > x)}$$

$$= \frac{\lim\limits_{h\to 0} \frac{1}{h}\left\{P(T_1 > x-h, T_2 > x+u, T_3 > x, T_4 > x) - P(T_1 > x, T_2 > x+u, T_3 > x, T_4 > x)\right\}}{\lim\limits_{h\to 0} \frac{1}{h}\left\{P(T_1 > x-h, T_2 > x, T_3 > x, T_4 > x) - P(T_1 > x, T_2 > x, T_3 > x, T_4 > x)\right\}}$$

$$= \frac{\lim\limits_{h\to 0} \frac{1}{h}\left\{C(S_1(x-h), S_2(x+u), S_3(x), S_4(x)) - C(S_1(x), S_2(x+u), S_3(x), S_4(x))\right\}}{\lim\limits_{h\to 0} \frac{1}{h}\left\{C(S_1(x-h), S_2(x), S_3(x), S_4(x)) - C(S_1(x), S_2(x), S_3(x), S_4(x))\right\}}$$

$$= \frac{C^{(1,0,0,0)}(S_1(x), S_2(x+u), S_3(x), S_4(x)) f_1(x)}{C^{(1,0,0,0)}(S_1(x), S_2(x), S_3(x), S_4(x)) f_1(x)} = \frac{C^{(1,0,0,0)}(S_1(x), S_2(x+u), S_3(x), S_4(x))}{C^{(1,0,0,0)}(S_1(x), S_2(x), S_3(x), S_4(x))}$$

with $C^{(1,0,0,0)}(u_1, u_2, u_3, u_4) = \frac{\partial}{\partial u_1} C(u_1, u_2, u_3, u_4)$. Similarly, the conditional probability for udder quarter 3 (T_3) is given by

$$P(T_3 > x+u | T_1 = x, T_2 > x, T_3 > x, T_4 > x) = \frac{C^{(1,0,0,0)}(S_1(x), S_2(x), S_3(x+u), S_4(x))}{C^{(1,0,0,0)}(S_1(x), S_2(x), S_3(x), S_4(x))}.$$

By taking the copula function of Model 2, we get the expressions (3) and (4).