Multivariate Survival Models for Interval-censored Udder Quarter Infection Times

K. Goethals

Thesis for submission in fulfilment of the requirements for the degree of
Doctor in Veterinary Sciences (PhD), Ghent University, 2011

Promotor: Prof. Dr. L. Duchateau
Co-promotor: Prof. Dr. P. Janssen

Faculty of Veterinary Medicine
Department of Comparative Physiology and Biometrics

The cover of the doctoral thesis was designed by Steven Delrue.

Multivariate Survival Models for Interval-censored Udder Quarter Infection Times

ISBN

Contents

List of abbreviations

AFT	accelerated failure time
$\boldsymbol{\beta}$	vector of regression parameters
C	copula
δ	censoring indicator
$f(.)$	density function
$F(.)$	cumulative distribution function
FL	front left
FR	front right
γ	shape parameter of the Weibull distribution
$h(.)$	hazard function
$H(.)$	cumulative hazard function
k	number of clusters
$\mathcal{L}(.)$	Laplace transform
L	likelihood
l	lower bound
λ	scale parameter of the Weibull distribution
n	sample size
RL	rear left
RR	rear right
RX	radiography
ρ	correlation between the frailties
$S(.)$	survival function
T	random variable representing the event time
θ	variance of the frailties in the frailty model
	copula function parameter in the copula model
τ	Kendall's τ
$\boldsymbol{x}$	vector of covariates
u	upper bound
US	ultrasound
Z	frailty

Chapter 1

An introduction to survival data and its analysis

1.1 Introduction

This introductory chapter describes some basic concepts in survival analysis. It contains notation and basic results on which the methodology developed in this thesis is based. Section 1.2 introduces the concept of survival data and describes what distinguishes survival analysis from other statistical fields. Section 1.3 describes basic quantities of survival analysis in a univariate setting. In Section 1.4 the concept of censoring is explained and Section 1.5 reviews models used to analyze univariate survival data. Section 1.6 introduces the concept of multivariate survival data, the main topic of this thesis. The frailty model is introduced in Section 1.7. Possible frailty distribution and estimation methods for the frailty model are discussed. The copula model is introduced in Section 1.8. Finally, Section 1.9 describes the examples used in this thesis to illustrate the developed methodology.

1.2 Survival data

Survival data, failure time data, lifetime data or time-to-event data are different names to describe data that deal with the time to an event. This event may be death (literally survival data), but also other events such as pregnancy, the recurrence of symptoms, the recovery from an illness or the occurrence of an infection are possible endpoints. Time-to-event data do not only arise in the field of demography, medicine or epidemiology; many other disciplines such as economics, engineering or sociology have to deal with time-to-event data. Consider, for instance, the time to leaving unemployment in economy (Nickell, 1979), the time to failure of a mechanical component of a machine in engineering (Lanternier et al., 2008) or the time to first use of marijuana in sociology (Turnbull and Weiss, 1978). In this thesis data sets from veterinary medicine will be used and focus will be on the time to infection of a cow udder quarter with a bacterium (Laevens et al., 1997).
A specific feature of survival data that distinguishes them from other data, is the presence of censoring. For a censored subject, the event time itself has not been observed; it is only known to fall into a certain interval (with possibly 0 as lower limit and/or ∞ as upper limit). Contrary to missing observations, censored observations still provide some information on the variable of interest (for more details on censoring, see Section 1.4).

1.3 Basic functions in survival analysis

Let $f(t)$ denote the density function of T, an absolutely continuous, non-negative random variable representing the event time of interest and $F(t) = P(T < t) = \int_0^t f(s)ds$ the corresponding cumulative distribution function. Let $S(t)$ denote the nonincreasing survival function defined as the probability that T exceeds a value t (Kaplan and Meier, 1958):

$$S(t) = 1 - F(t) = P(T \geqslant t) = \int_t^\infty f(s)ds.$$

An important concept in survival analysis is the hazard function $h(t)$ defined as

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}.$$

It represents the instantaneous failure rate at time t, given that the subject does not fail until time t. $h(t)\Delta t$ may be viewed as the approximate probability of a subject failing in the next instant.
The survival, density and hazard functions have the following one-to-one relationships:

$$\begin{aligned} f(t) &= -\frac{dS(t)}{dt} \\ h(t) &= \frac{f(t)}{S(t)} = \frac{-d \log S(t)}{dt} \\ S(t) &= \exp\left(-H(t)\right) \end{aligned}$$

where $H(t) = \int_0^t h(s)ds$ is the cumulative hazard function.
The Laplace transform of a random variable T is also an important concept in survival analysis:

$$\mathcal{L}(s) = E[\exp(-sT)] = \int_0^\infty \exp(-st)f(t)dt. \tag{1.1}$$

1.4 Censoring and truncation

1.4.1 Left, right and interval censoring

As mentioned above, the presence of censoring is a specific feature that makes survival analysis different from other statistical disciplines. Depending on the situation different types of censoring exist: an observation is right-censored if the actual (unobserved) event time is larger than the (observed)

censoring time. A typical situation that yields right-censored observations is one in which the study has to end at a predefined point in time, for example due to time constraints or resource limitations. Other possible reasons for right censoring are drop out or loss to follow up of subjects (because they moved or do not show up at planned visits). On the other hand, an observation is left-censored when the actual (unobserved) event time is smaller than the (observed) censoring time, i.e., when the event of interest has already occurred before the subject is observed in the study. Interval-censored data arise when the exact time-to-event is not known; it is only known that the event occurred within a certain interval of time. For instance, when a subject is not monitored continuously, but only examined at scheduled visiting times, we only know that the event of interest happened between the last visit at which the event had not taken place yet and the first visit at which the event has taken place.

Formally, let T be the real, possibly unobserved, event time. If an observation is interval-censored, let L and U denote the lower and upper bound of the interval, respectively, then $L < T \leqslant U$. Left- and right-censored observations can be considered special cases of interval-censored observations. For a left-censored observation L is the start of the at risk time (time 0) and U is the censoring time. Similarly $U = \infty$ and L is the censoring time (for example the end of the study or the last visiting time the subject was seen before the end of the study) for right-censored observations. When the event time is known exactly, $L = U$. Intervals can be recorded as open, half open or closed; if T is continuous, they represent the same observed information about T. The recording of closed intervals allows for exact observations. To cover all situations we will use the notation $\lfloor L, U \rceil$.

Figures 1.1 and 1.2 present the different types of censoring. In Figure 1.1 observations are either exact or right-censored. Consider, for example, a study in which time to culling in heifers is studied for an entire lactation period (roughly 300-350 days, different for every cow) (De Vliegher et al., 2005). Each cow enters the study at the first day of its lactation (time 0) and if the cow is culled during the lactation the exact culling day is known (cow 1 in Figure 1.1 for example is culled at day 200 of its lactation period). The observations for cows 2 and 3 are right-censored: the observation for cow 2 because it was not culled before the end of its lactation period (its censoring time is 350 days), the observation for cow 3 because it was perhaps sold at day 100 of its lactation period. In Figure 1.2 observations are either interval-, left- or right-censored. For an example we refer to the mastitis data (Section 1.9.2), a study on intramammary infections in cows. Each cow enters the study at the first day of its lactation (time 0) and is monthly

screened for bacterial infections at the udder quarter level. In Figure 1.2 we consider the left front udder quarter of a few cows. The observation for the udder quarter of cow 1 is interval-censored: the udder quarter got infected between the visit of October and the visit of November. The observations for the udder quarters of cows 2 and 3 are right-censored: the udder quarter of cow 2 was not infected before the end of the study, cow 3 again could have been sold or could have been culled. The end of the study is the censoring time for cow 2. The censoring time for cow 3 is the last visiting time at which the cow was still on the farm. The udder quarter of cow 4 was already infected before the first visiting time in June and is thus left-censored. The first visiting time is its censoring time.

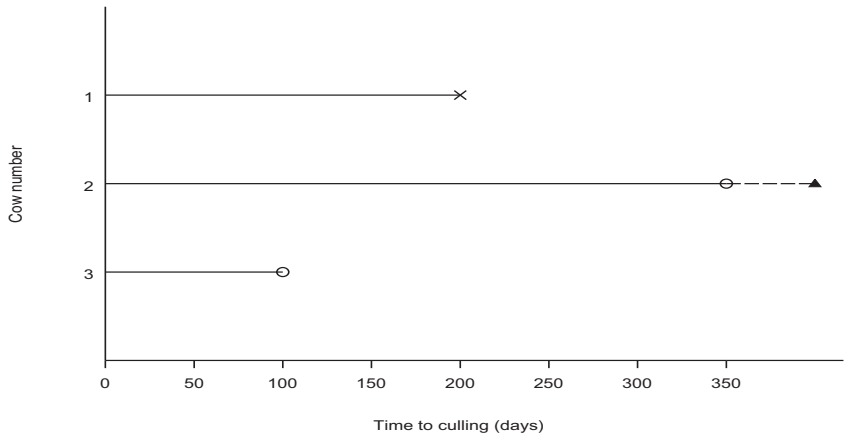

Figure 1.1: Exact and right-censored observations. An 'x' denotes an observed event, an 'o' a censored observation, an '▲' an unobserved event.

Different types of right and interval censoring exist (Klein and Moeschberger, 2003). The censoring mechanism that stops the study at the same fixed time point for all subjects is called Type I right censoring. In the case of Type II right censoring the study stops if a prespecified number of subjects has experienced the event. Sometimes the event of interest can not be observed because the subject is removed from the study. This is termed random censoring. Two possible causes of random censoring are accidental deaths and subject withdrawal from the study. In some studies, the censoring scheme is a combination of random censoring and type I censoring. In such studies

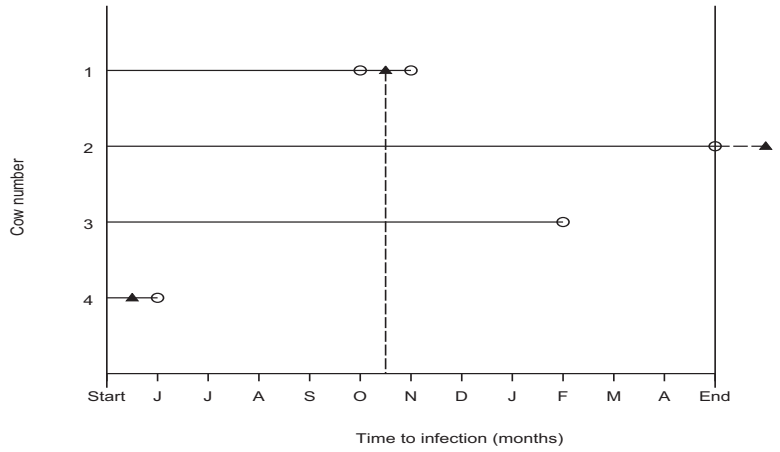

Figure 1.2: Left-, right- and interval-censored observations. An '▲' denotes an unobserved event, an 'o' a censored observation.

some subjects are randomly censored, for example due to drop out, whereas others are type I censored when the fixed study period ends. The mastitis study, introduced in Section 1.9.2, is an example of a study with a combination of randomly censored and type I censored observations.

Case I interval-censored data or current status data occur when there is only a (fixed) censoring time and it is only known that the event occurred either before or after the censoring time. Therefore, each observation time is either left- or right-censored. Case I interval-censored data are for example encountered in tumorigenicity experiments where an animal needs to be sacrificed to determine whether a (nonlethal) tumor is present or absent. The time to tumor onset is not observed directly, it is only known to be less than or greater than the time of sacrifice.

Case II interval censoring refers to the situation with two (fixed) censoring times (L, U), where $L < U$ and $L, U \in (0, \infty)$, and the available information restricts to whether the event time T is either smaller than L, between L and U or larger than U.

The extension of case II interval-censored data to the situation in which there are k (fixed) censoring times is called case k interval censoring (Sun, 2006).

Grouped event time data (Lawless, 2003) occur if the predetermined visiting

schedule is followed exactly by all study subjects. But often, subjects miss certain appointments and return with a changed status or visit the clinic at times that are convenient to them instead of at predetermined observation times.
When there are two related events in a survival study, say V and Y, $V \leqslant Y$, and $T = Y - V$ is the event time of interest, observations on both V and Y can be right- or interval-censored. This type of data is called doubly censored survival data (De Gruttola and Lagakos, 1989; Sun et al., 2004) and is often seen in disease progression studies where the two events may represent infection and the onset of a disease, respectively. In AIDS studies, for example, observations are often doubly censored with V the time of HIV infection, determined by periodic blood tests and therefore interval-censored, and Y the onset of AIDS, often determined by the CD4 cell count in blood, tested periodically and thus also interval-censored.

1.4.2 Noninformative censoring

We will assume throughout this thesis noninformative censoring, a commonly used assumption in survival analysis. This means that the censoring procedure does not contain any information on the parameters used to model the event time. Without this assumption statistical inference is much harder.
Right-censored data consist of either exact event times or right-censored observations. The likelihood for a sample of size n is then given by (Klein and Moeschberger, 2003)

$$L = \prod_{i=1}^{n} \left[(1 - G(y_i))\, f(y_i) \right]^{\delta_i} \left[(1 - F(y_i))\, g(y_i) \right]^{1-\delta_i},$$

with $f(.)$ the density function of the event times with corresponding cumulative distribution function $F(.)$, $g(.)$ the density function of the censoring times with corresponding cumulative distribution function $G(.)$ and δ_i the censoring indicator, taking the value one if the event has been observed, otherwise δ_i takes the value zero. Under the assumption of noninformative censoring the factors $(1 - G(y_i))$ and $g(y_i)$ are not informative for inference on the survival function and can therefore be deleted from the likelihood, resulting in the following simplified likelihood:

$$L \approx \prod_{i=1}^{n} \left(f(y_i) \right)^{\delta_i} \left(S(y_i) \right)^{1-\delta_i}.$$

The noninformative censoring condition in case of interval-censored data can be formally defined in different ways. Self and Grossman (1986) proposed the following definition

$$dF_{T|L,U}(t|l,u) = \frac{dF_T(t)}{P(T \in \lfloor l,u \rfloor)} 1_{\{t:t\in\lfloor l,u\rfloor\}}(t),$$

with 1 the indicator function, equal to one if the condition is true and zero otherwise. This definition states that the only information provided by the censoring interval $\lfloor l,u \rfloor$ about the event time is that the interval contains t. Equivalent definitions are given by Gomez et al. (2004)

$$dF_{L,U|T}(l,u|t) = \frac{dF_{L,U}(l,u)}{P(T \in \lfloor l,u \rfloor)} 1_{\{(l,u):t\in\lfloor l,u\rfloor\}}(l,u)$$

i.e., the observables (l,u) are not influenced by the specific value of T $\in \lfloor l,u \rfloor$, and by Heitjan and Rubin (1991)

$$dF_{L,U|T}(l,u|t) = dF_{L,U|T}(l,u|t') \text{ on } \left\{(l,u) : t \in \lfloor l,u \rfloor \text{ and } t' \in \lfloor l,u \rfloor\right\}$$

i.e., two specific values of T that are consistent with the observables always provide the same information.
Therefore, the contribution of an individual with observed interval $\lfloor l,u \rfloor$ to the likelihood, $dF_{L,U}(l,u) = P(L \in dl, U \in du, T \in \lfloor l,u \rfloor)$, can be simplified to $P(l < T \leqslant u) = S(l) - S(u)$. Oller et al. (2004) showed that the three definitions above are equivalent and justify the use of the simplified likelihood

$$L \approx \prod_{i=1}^{n} S(l_i) - S(u_i)$$

for interval-censored data.
In the mastitis study the visits made by the veterinarian to the herds were planned in advance, independently of the event of interest (i.e. infection), therefore, we assume that the censoring mechanism is noninformative. However, if the veterinarian was asked to visit the herd for a particular cow showing clinical signs of infection, the noninformative censoring condition would not be valid because the event of interest, the infection, induced the visit.

1.5 Regression models in survival analysis

1.5.1 The proportional hazards model

The most popular regression model for right-censored survival data, especially in the field of medicine and biostatistics, is the proportional hazards

model (Cox, 1972). For a given vector of covariates $\boldsymbol{x}$, the hazard function $h(t)$ is expressed as the product of an unspecified baseline hazard function $h_0(t)$ and a positive function of the covariates, usually the exponential of a linear function of $\boldsymbol{x}$:

$$h(t) = h_0(t) \exp(\boldsymbol{x}^t \boldsymbol{\beta}),$$

with $\boldsymbol{x}^t$ the transpose of the vector $\boldsymbol{x}$ and $\boldsymbol{\beta}$ the vector of regression parameters. Therefore, the model assumes a common baseline hazard for all subjects in the study population and specifies that the exponentially transformed covariates act multiplicatively on the hazard function.
An important feature of the proportional hazards model is the separation of the time effect in the baseline hazard and the covariate effect in the exponential function. Therefore, the ratio of the hazard functions for two subjects with different covariate information is constant over time. Consider for example the simple two-sample situation where x can be 0 or 1 depending on the group the subject belongs to. Then the hazard ratio is equal to

$$\frac{h_0(t) \exp \beta}{h_0(t)} = \exp \beta.$$

The baseline hazard function $h_0(t)$ can be assumed to have a particular parametric form or can be left unspecified. A popular choice for the parametric baseline hazard is

$$h_0(t) = \lambda \gamma t^{\gamma - 1} \qquad \text{with } \lambda > 0, \gamma > 0,$$

leading to event times that are Weibull distributed with shape parameter γ and scale parameter λ (Weibull, 1951). The Weibull distribution is a popular choice for the event times in survival analysis because it is a fairly flexible distribution that describes the evolution of the hazard well in practice. For $\gamma < 1$ the hazard decreases monotonically over time, for $\gamma > 1$ the hazard is monotone increasing and for $\gamma = 1$ the hazard is constant over time, corresponding to exponentially distributed event times. Figure 1.3 shows Weibull hazard functions with scale parameter equal to 0.9 and different shape parameters. Other possible choices for the distribution of the event times include the exponential and Gompertz distribution (Klein and Moeschberger, 2003). The form of the baseline hazard $h_0(t)$ can also be left unspecified, but the effect of the covariates on the hazard function still needs to be modeled parametrically. Since the model then contains a parametric factor $\boldsymbol{x}^t \boldsymbol{\beta}$ and a nonparametric baseline hazard, it is called semiparametric. One of the main reasons for the popularity of the semiparametric Cox

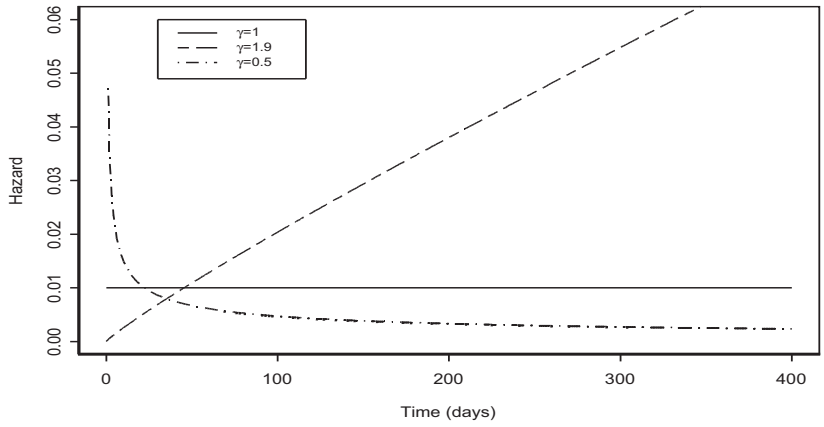

Figure 1.3: Weibull hazard functions with different shape parameters

proportional hazards model is the existence of a simple and efficient inference procedure for the regression parameters in case of right-censored data, namely the partial likelihood maximization procedure introduced by Cox (1972, 1975). This likelihood no longer contains the baseline hazard $h_0(t)$, but is only a function of the unknown regression parameters $\boldsymbol{\beta}$, which can be estimated through maximization of the likelihood. Asymptotic properties for the partial likelihood estimator for $\boldsymbol{\beta}$ are well established (Gill, 1984). Furthermore, the technique is implemented in most statistical software packages, which further explains its popularity.

1.5.2 The accelerated failure time model

A useful, however less frequently used, alternative to the proportional hazards model is the accelerated failure time (AFT) model (Lawless, 2003). The AFT model is best described in terms of the survival function:

$$S(t) = S_0(\Phi t),$$

with $S_0(.)$ the baseline survival function and Φ ($\Phi > 0$) the acceleration factor, often put equal to $\exp(\boldsymbol{x}^t\boldsymbol{\beta})$. If we think in terms of a treated and a control group (the baseline), interpretation is as follows: the percentage of subjects in the treatment group that lives longer than t equals the percentage

of subjects in the control group that lives longer than Φt. Values of Φ below one are in favour of the treatment, as the event time is then prolonged under treatment. Values of Φ larger than one are in favour of the control. The effect of covariates on the hazard function is given by

$$h(t) = \exp(\boldsymbol{x}^t\boldsymbol{\beta})h_0\left(\exp(\boldsymbol{x}^t\boldsymbol{\beta})t\right).$$

Again, a parametric form for the baseline hazard can be assumed (Klein and Moeschberger, 2003) and parameters can be estimated using the method of maximum likelihood. Techniques to fit parametric AFT models are available in most commercial statistical software packages. Semiparametric procedures however are computationally demanding and not implemented in commonly used statistical software packages. This probably caused the proportional hazards model to be more popular in practice.
However, in the AFT model the regression parameter indicates an increasing or decreasing effect on time in a direct way, a concept that is more easily understandable for non-statisticians, whereas in the proportional hazards model, the effect is on the hazard, which is probably harder to understand for practitioners. Another advantage of the AFT model is that the parameter estimates of the covariates included in the model do not change when other, important, covariates are omitted. This property does not hold for the proportional hazards model (Hougaard, 1999). A more detailed comparison of the proportional hazards model and the AFT model can be found in Orbe et al. (2002). In this thesis most models are based on the fully parametric proportional hazards formulation.

1.5.3 The loglinear model representation

Most statistical software packages, e.g. SPlus, R and SAS, do not report the parameter estimates for the parametric proportional hazards model or the accelerated failure time model. Instead they supply parameter estimates for the loglinear model. Therefore, we will introduce this model here and show the relationships between the parameter estimates of the loglinear model and the proportional hazards and accelerated failure time model in the case of Weibull distributed event times.
In the loglinear model the event time is modeled directly, instead of modeling the hazard function (proportional hazards model) or the survival function (accelerated failure time model)

$$\log(T) = \mu + \boldsymbol{x}^t\boldsymbol{\alpha} + \sigma E,$$

with T the event time, μ the intercept, $\boldsymbol{x}$ the vector of covariates, $\boldsymbol{\alpha}$ the vector of regression parameters, σ the scale parameter and E the random error term. If a Gumbel distribution is assumed for the error term E

$$E \sim f_E(e) = \exp\left(e - \exp(e)\right) \quad \text{for } -\infty < e < \infty$$

the event times follow a Weibull distribution.
The survival function in case of Weibull distributed event times in the loglinear model is

$$S(t) = \exp\left[-\exp\left(-\mu/\sigma\right) t^{1/\sigma} \exp\left(\boldsymbol{x}^t\left(-\boldsymbol{\alpha}/\sigma\right)\right)\right]. \tag{1.2}$$

For the Weibull accelerated failure time model the survival function can be written as

$$S(t) = \exp\left(-\lambda t^{\gamma} \exp\left(\gamma \boldsymbol{x}^t \boldsymbol{\beta}\right)\right). \tag{1.3}$$

Comparing (1.2) and (1.3) it can be seen that the following relationships hold

$$\lambda = \exp(-\mu/\sigma) \quad \gamma = \sigma^{-1} \quad \boldsymbol{\beta} = -\boldsymbol{\alpha}.$$

The survival function in the Weibull proportional hazards model is given by

$$S(t) = \exp\left(-\lambda t^{\gamma} \exp\left(\boldsymbol{x}^t \boldsymbol{\beta}\right)\right). \tag{1.4}$$

Comparing (1.2) and (1.4) it can be seen that the two models correspond with

$$\lambda = \exp(-\mu/\sigma) \quad \gamma = \sigma^{-1} \quad \boldsymbol{\beta} = -\boldsymbol{\alpha}/\sigma. \tag{1.5}$$

Therefore, the parameter estimates from the loglinear model can easily be transformed into parameter estimates for either the Weibull accelerated failure time model or the Weibull proportional hazards model.
Obtaining variance estimates for the parameters of the Weibull accelerated failure time model and the Weibull proportional hazards model based on variance estimates from the loglinear model is less straightforward. Approximations can be obtained using the delta method (Oehlert, 1992). Variance estimates (or approximations) for the parameters from the Weibull accelerated failure time model are given by:

$$\begin{aligned}
\mathrm{Var}(\hat{\lambda}) &\approx \exp(-\hat{\mu}/\hat{\sigma})^2 \hat{\sigma}^{-2} \mathrm{var}(\hat{\mu}) + \exp(-\hat{\mu}/\hat{\sigma})^2 \hat{\mu}^2 \hat{\sigma}^{-4} \mathrm{var}(\hat{\sigma}) \\
&\quad -2\exp(-\hat{\mu}/\hat{\sigma})^2 \hat{\mu} \hat{\sigma}^{-3} \mathrm{cov}(\hat{\mu}, \hat{\sigma}) \\
\mathrm{Var}(\hat{\gamma}) &\approx \hat{\sigma}^{-4} \mathrm{var}(\hat{\sigma}) \\
\mathrm{Var}(\hat{\beta}_k) &= \mathrm{var}(\hat{\alpha}_k).
\end{aligned}$$

Variance approximations for the parameters from the Weibull proportional hazards model are given by

$$\begin{aligned} \mathrm{Var}(\hat{\lambda}) &\approx \exp(-\hat{\mu}/\hat{\sigma})^2\hat{\sigma}^{-2}\mathrm{var}(\hat{\mu}) + \exp(-\hat{\mu}/\hat{\sigma})^2\hat{\mu}^2\hat{\sigma}^{-4}\mathrm{var}(\hat{\sigma}) \\ &\quad -2\exp(-\hat{\mu}/\hat{\sigma})^2\hat{\mu}\hat{\sigma}^{-3}\mathrm{cov}(\hat{\mu},\hat{\sigma}) \qquad (1.6) \\ \mathrm{Var}(\hat{\gamma}) &\approx \hat{\sigma}^{-4}\mathrm{var}(\hat{\sigma}) \qquad (1.7) \\ \mathrm{Var}(\hat{\beta}_k) &\approx \hat{\sigma}^{-2}\mathrm{var}(\hat{\alpha}_k) + \hat{\alpha}_k^2\hat{\sigma}^{-4}\mathrm{var}(\hat{\sigma}) - 2\hat{\alpha}_k\hat{\sigma}^{-3}\mathrm{cov}(\hat{\alpha}_k,\hat{\sigma}), \qquad (1.8) \end{aligned}$$

with β_k and α_k the k^{th} component in the regression parameter vector of the Weibull accelerated failure time model or proportional hazards model and the loglinear model, respectively.

1.6 Multivariate survival data

So far, the basic concepts in survival analysis are introduced in a univariate setting. Classical survival analysis techniques assume that survival times of different subjects are independent. Although this assumption may be valid in many situations, it will be violated in others. Indeed, survival times are frequently not independent of each other because the subjects have some feature in common. For example, animals within a litter will be more alike than animals from different litters because of genetic and environmental influences. Such data are known as clustered or correlated survival data. In the main example of this thesis udder quarter infection times are clustered within the cow. In recent years, extensive research on clustered survival data has been carried out. Marginal models ignore the correlation between event times but provide consistent parameter estimates (Wei and Glidden, 1997). An appropriate version of the asymptotic variance-covariance matrix of the estimators which takes into account the clustering is also available. The frailty model (Duchateau and Janssen, 2008) models the correlation between event times by introducing a common random effect called frailty. The event times are independent conditionally on the frailty. In the copula model, the copula couples the marginal survival functions and the joint survival function and determines the type of correlation (Genest and MacKay, 1986). Contrary to the marginal model, the frailty model and the copula model provide a measure for the strength of the correlation between event times next to the estimated covariate effects. For a discussion on frailty models and copula models we refer to the next sections and Chapter 3. For a discussion on the marginal model we refer to Chapter 4.

1.7 The frailty model

In this section and the next section we will focus on two models that are widely used to fit multivariate survival data: the frailty model and the copula model. Both models provide an estimate of the correlation between event times in a cluster. The frailty model can also be used in a univariate setting to account for unobserved heterogeneity. In this section the univariate, shared and correlated frailty model are introduced and possible frailty distributions and estimation methods for the frailty model are described. In the next section we introduce the copula model and briefly describe the two-stage estimation approach for copulas.
In general, frailty is defined as susceptibility to a certain event. This susceptibility can be individual (the univariate frailty model) or can be (partly) shared by different members of a cluster (the shared frailty model, the correlated frailty model). In the different frailty models, the frailty term is a positive random variable following some distribution, for example gamma, positive stable or lognormal. The frailty proportional hazards model is a proportional hazards model in which the hazard of a subject depends on covariates and on an unobserved frailty term Z, which acts multiplicatively on the baseline hazard.
In the following sections different frailty models (univariate, shared and correlated frailty model), frailty distributions (the gamma, positive stable and lognormal distribution) and estimation methods (maximum likelihood, Expectation-Maximization (EM)-algorithm, penalized partial likelihood) for the shared frailty model will be described.

1.7.1 The univariate frailty model

Ordinary survival analysis assumes that the population under study is homogeneous, this means that the risk of experiencing the event of interest is the same for every subject in the population with the same covariate information. However, subjects can differ greatly among themselves due to covariates such as age, gender, length, socio-economic status, education level or housing (in animal studies). If covariates are known, they can be included in the analysis, but it is nearly always impossible to include all important covariates in the model. It may be impossible to measure the covariate due to financial or time constraints or the investigator might be unaware of its existence. Therefore, there is always variability in the hazard function in a study population. This variability not explained by observed covariates is called unobserved heterogeneity.

One way of dealing with unobserved heterogeneity is the use of univariate frailty models. In the univariate frailty model each study subject has its own frailty, and it is assumed that the more frail subjects will experience the event earlier than the lesser frail. The concept of frailty was first introduced in survival analysis by Beard (1959) to improve the modeling of mortality in a heterogeneous population. However, Beard (1959) used the term longevity factor rather than frailty. The actual term *frailty* was introduced by Vaupel et al. (1979).
In the univariate frailty model the variance of the frailties, θ, determines the degree of heterogeneity between subjects in the study population: the larger the variance, the more heterogeneity in the population.
For a total of n subjects, the univariate frailty model is given by

$$h_j(t) = h_0(t) z_j \exp(\boldsymbol{x}_j^t \boldsymbol{\beta}), \qquad j = 1, \ldots, n$$

with $h_j(t)$ the hazard for subject j, $h_0(t)$ the baseline hazard, $\boldsymbol{x}_j$ the vector of covariates for the j^{th} subject, $\boldsymbol{\beta}$ the vector of regression parameters and z_j the frailty for the j^{th} subject, coming from a density $f_Z(z)$. The correlation between the frailties z_j is equal to zero.

1.7.2 The shared frailty model

Statistical models for time-to-event data, including the Cox proportional hazards model, implicitly assume that observations are statistically independent. This assumption does not hold in all situations. In the mastitis data set given in Section 1.9.2, for example, the infection times of the four udder quarters are obviously correlated within cow and also the two diagnosis times (RX and US) in the time to diagnosis data set, introduced in Section 1.9.1, are not independent.
Shared frailty models were developed by Clayton (1978) to deal with this type of data, called correlated or multivariate survival data. In the shared frailty model subjects in the same cluster share the same frailty term, hence the name shared frailty model.
Since the frailty Z is common to all subjects in a cluster, it is responsible for creating correlation between the event times in a cluster. This correlation is always positive and the type of correlation is determined by the choice of the frailty distribution. The variance of the frailties is a measure for the heterogeneity between clusters. Given the frailties, observations in a cluster are independent.
Assume we have a total of n subjects coming from k different clusters, cluster i, $i = 1, \ldots k$, having n_i subjects ($n = \sum_{i=1}^{k} n_i$). The shared frailty model

is given by

$$h_{ij}(t) = h_0(t) z_i \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}),$$

with $h_{ij}(t)$ the conditional hazard function at time t for the j^{th} subject of the i^{th} cluster, $j = 1, \ldots, n_i, i = 1, \ldots k$, $h_0(t)$ the baseline hazard, $\boldsymbol{x}_{ij}$ the vector of covariates for the j^{th} subject of the i^{th} cluster, $\boldsymbol{\beta}$ the vector of regression parameters and z_i the frailty for the i^{th} cluster, coming from a density $f_Z(z)$. The correlation between the frailties of different clusters is equal to zero. The correlation between individual frailties of the members in a cluster is equal to one.

1.7.3 The correlated frailty model

Another approach to model multivariate survival data is the correlated frailty model. In the correlated frailty model each subject in a cluster has its specific frailty term. The individual frailties of the members of a cluster are correlated, thereby inducing correlation between the event times in a cluster (Yashin et al., 1995). The shared frailty model and the univariate frailty model can be interpreted as special cases of the correlated frailty model where the correlation between the frailties in a cluster is equal to one and zero, respectively.
Assume again a total of n subjects coming from k different clusters, the correlated frailty model is given by

$$h_{ij}(t) = h_0(t) z_{ij} \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}),$$

with $h_{ij}(t)$, $h_0(t)$, $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}$ defined as in Section 1.7.2 and z_{ij} the frailty term for the j^{th} subject of the i^{th} cluster, coming from a density $f_Z(z)$. The correlation between individual frailties of the members in a cluster is equal to ρ. The correlation between frailties of different clusters is equal to zero.

1.7.4 Frailty distributions

As mentioned in the previous section different distributions have been proposed for the frailty term. In this section we will discuss the gamma distribution, the lognormal distribution and the positive stable distribution in more detail. Both the gamma and the positive stable distribution belong to the three parameter family of power variance function distributions, introduced by Hougaard (1986b).

The gamma distribution

The most common choice for the frailty distribution is the one-parameter gamma distribution gamma$(1/\theta, 1/\theta)$ (Vaupel et al., 1979). The density is given by

$$f_Z(z) = \frac{z^{1/\theta-1}\exp(-z/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}, \tag{1.9}$$

with $\theta > 0$, $\Gamma(1/\theta) = \int_0^\infty t^{1/\theta-1}\exp(-t)dt$, $\mathrm{E}(Z) = 1$ and $\mathrm{Var}(Z) = \theta$. The Laplace transform is

$$\mathcal{L}(s) = (1+\theta s)^{-1/\theta}. \tag{1.10}$$

The popularity of the gamma distribution as a frailty distribution is based on mathematical and computational aspects: the gamma distribution has a simple distribution and simple Laplace transform. This will make inference less complicated. Most importantly, assuming a gamma distribution for the frailty enables us to integrate out the frailties from the conditional likelihood, resulting in a simple and closed form expression for the marginal likelihood which can then be maximized to obtain parameter estimates. For more details, see Section 1.7.5. In a similar way simple, closed form expressions for the marginal survival and hazard function can be derived. Unfortunately, there are no biological reasons that justify the choice of a gamma distribution for the frailty variable, however arguments for the use of the gamma distribution for frailties in duration analysis are given by Abbring and Van Den Berg (2007).

The lognormal distribution

In practice, the gamma distribution and the lognormal distribution are most often used to model the frailty term and in most software packages only those two options are available as frailty distribution. The use of the lognormal distribution for the frailty term originates from the mixed model framework, where a normal distribution is assumed for the random effect W. McGilchrist (1993) proposes the following model

$$h(t) = h_0(t)\exp(\boldsymbol{x}^t\boldsymbol{\beta} + w),$$

the frailty is then $Z = \exp(W)$. Two variants of the lognormal model can be used. One idea is to assume a normal density with $\mathrm{E}(W) = 0$ and $\mathrm{Var}(W) = \sigma^2$ for the random effect W. The corresponding density function of Z is then the following lognormal density

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\log z)^2}{2\sigma^2}\right).$$

The mean and variance of the frailty are then given by

$$\begin{aligned} \mathrm{E}(Z) &= \exp\left(\sigma^2/2\right) \\ \mathrm{Var}(Z) &= \exp(2\sigma^2) - \exp(\sigma^2). \end{aligned}$$

The other option is to restrict the expected value of the frailty Z to one, however, this approach is not often used.
There is no simple expression for the Laplace transform.

The positive stable distribution

The positive stable distribution was introduced as a frailty distribution by Hougaard (1986a). The distribution is given by

$$f_Z(z) = -\frac{1}{\pi z} \sum_{k=1}^{\infty} \frac{\Gamma(k\theta + 1)}{k!} \left(-z^{-\theta}\right)^k \sin(\theta k \pi), \tag{1.11}$$

with $0 \leqslant \theta < 1$. This distribution has infinite mean and the variance is therefore also undetermined. Despite the fact that the form of the distribution is complicated, the Laplace transform has the simple form

$$\mathcal{L}(s) = \exp\left(-s^{\theta}\right). \tag{1.12}$$

The most interesting feature of the positive stable distribution is the fact that it is the only frailty distribution that preserves the proportional hazards assumption in the marginal hazards after integrating out the frailty.

Diagnostics

Few results are available on comparing models with different frailty distributions. Most research on diagnostic tests for the frailty distribution has been undertaken for the bivariate gamma frailty model.
Oakes (1982, 1989) discuss a diagnostic test for the gamma frailty distribution for bivariate data without censoring based on the cross ratio function. The cross ratio function is the ratio of the hazard of the first subject of a pair experiencing the event at t_1, given that the second subject experiences the event at t_2 over the hazard that the first subject experiences the event at t_1, given that the second subject has not experienced the event yet at t_2. For the gamma distribution the cross ratio function is constant and equal

to $\theta + 1$. Oakes (1982, 1989) propose a nonparametric estimate of the cross ration function for bivariate data without censoring based on an observable risk set. If the gamma distribution is an appropriate frailty distribution the nonparametric estimate should be equal to $\theta + 1$. The technique can also be applied to check the validity of the assumption of a positive stable distribution as the frailty distribution. For the positive stable distribution the cross ratio function $\psi(t_1, t_2)$ is given by

$$\psi(t_1, t_2) = 1 + \frac{\theta - 1}{\theta \log S_f(t_1, t_2)}.$$

Shih and Louis (1995a) propose a diagnostic test for the multivariate gamma frailty model based on the evolution of the conditional posterior mean of the frailties over time. The average of the posterior frailty means should take constant value one for all timepoints if the gamma distribution assumption for the frailty is correct.
Other diagnostic techniques to evaluate the frailty distribution assumption can be found in Glidden (1999); Cui and Sun (2004); Economou and Caroni (2005).

1.7.5 Estimation methods for the frailty model

In this section we will discuss estimation methods for the univariate or shared frailty model. They will be presented in the context of shared frailty models. For a discussion on estimation methods for the correlated frailty model, we refer to Chapter 6.
Parameter estimates in the frailty model are obtained by maximizing the marginal likelihood. The marginal likelihood is obtained by either integrating out the frailties directly from the conditional likelihood or it is based on the derivatives of the marginal survival function, which can be obtained by integrating out the frailties from the conditional survival function using the Laplace transform (see (1.1)). The former approach is usually used in the gamma frailty model and the lognormal frailty model while the latter is used in the positive stable model because the form of the positive stable distribution is complicated but its Laplace transform is simple.
We first describe in general how the marginal survival function can be obtained from the conditional survival function using the Laplace transform. The joint conditional survival function is given by

$$S_i(\boldsymbol{t}_{n_i}) = \exp\left[-z_i\left(H_0(t_{i1})\exp\left(\boldsymbol{x}_{i1}^t\boldsymbol{\beta}\right) + \ldots + H_0(t_{in_i})\exp\left(\boldsymbol{x}_{in_i}^t\boldsymbol{\beta}\right)\right)\right],$$

with $\boldsymbol{t}_{n_i} = (t_1, \ldots, t_{n_i})$, n_i the number of members in a cluster and $H_0(t) = \int_0^t h_0(s)ds$ the cumulative baseline hazard. The joint marginal survival function is obtained from the joint conditional survival function by integrating out the frailty with respect to the frailty distribution

$$\begin{aligned} S(\boldsymbol{t}_{n_i}) &= \int_0^\infty \exp\left(-z\sum_{j=1}^{n_i} H(t_{ij})\right) f_Z(z)dz \\ &= \mathrm{E}\left[\exp\left(-Z\sum_{j=1}^{n_i} H(t_{ij})\right)\right], \end{aligned} \tag{1.13}$$

with $H(t_{ij}) = H_0(t_{ij})\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)$ the cumulative hazard function for the j^{th} subject of the i^{th} cluster. The last line of equation (1.13) is the Laplace transform of Z, $\mathcal{L}(s) = \mathrm{E}\left[\exp\left(-Zs\right)\right]$, evaluated at $s = \sum_{j=1}^{n_i} H(t_{ij})$. Therefore, we can write

$$S(\boldsymbol{t}_{n_i}) = \mathcal{L}\left(\sum_{j=1}^{n_i} H(t_{ij})\right). \tag{1.14}$$

We now discuss the marginal likelihood for the gamma, positive stable and lognormal frailty model in detail.

The gamma frailty model

In the gamma frailty model a closed form expression for the marginal likelihood of cluster i, $i = 1, \ldots, k$, can be obtained by integrating out the frailties from the conditional likelihood:

$$\begin{aligned} L_{marg,i}(\boldsymbol{\zeta}) &= \int_0^\infty \prod_{j=1}^{n_i} (h_0(y_{ij})z\exp(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}))^{\delta_{ij}} \exp(-H_0(y_{ij})z\exp(\boldsymbol{x}_{ij}^t\boldsymbol{\beta})) \\ &\quad \times \frac{z^{1/\theta-1}}{\theta^{1/\theta}\Gamma(1/\theta)} \exp\left(-z/\theta\right) dz, \end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard. The marginal loglikelihood is then given by (Klein, 1992)

$$\begin{aligned} l_{marg}(\boldsymbol{\zeta}) &= \sum_{i=1}^{k} \Bigg[d_i \log\theta - \log\Gamma(1/\theta) + \log\Gamma(1/\theta + d_i) \\ &\quad -(1/\theta + d_i) \log\left(1 + \theta \sum_{j=1}^{n_i} H_0(y_{ij}) \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta})\right) \\ &\quad + \sum_{j=1}^{n_i} \delta_{ij} \left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \log h_0(y_{ij})\right) \Bigg], \end{aligned}$$

with $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ the number of observed events in cluster i.

If a parametric assumption is made for the baseline hazard, the marginal likelihood is fully parametric and classical maximum likelihood techniques can be used to estimate the parameters. Standard errors can be obtained from the inverse of the observed information matrix.

If the baseline hazard is left unspecified, the model is semiparametric and direct maximization of the marginal likelihood is no longer possible. A combination of partial likelihood ideas and the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) can then be used to obtain parameter estimates (Klein, 1992). The EM algorithm iterates between an expectation and maximization step until convergence. In the expectation step, the expected values of the unobserved frailties conditional on the observed information and the current parameter estimates are obtained. In the maximization step, these expected values are considered to be fixed and new estimates of the parameters of interest are obtained by maximization of the likelihood, given the expected values.

Another way to obtain parameter estimates in the semiparametric gamma frailty model is the penalized partial likelihood approach. This approach is based on the observation that the likelihood consists of two parts. The first part consists of the likelihood of the data given the frailties and can be transformed into a partial likelihood expression. The second part corresponds to the distribution of the frailties, and is considered to be a penalty term, i.e., frailties far away from the mean 1 contribute a large penalty to the likelihood. More details on the EM-algorithm and the penalized partial likelihood approach for the gamma frailty model can be found in Duchateau et al. (2002).

The lognormal frailty model

Since the Laplace transform of the lognormal density is intractable, frailties can not be integrated out analytically and a closed form expression for the marginal likelihood does not exist. In a parametric setting numerical integration will be needed to integrate out the frailties to obtain the marginal likelihood which can then be maximized.
For the semiparametric lognormal frailty model the penalized partial likelihood approach can be used to estimate the parameters (McGilchrist and Aisbett, 1991; McGilchrist, 1993).

The positive stable frailty model

Although the marginal likelihood expression with the positive stable density has a closed form, it is much more complex then the one for the gamma density. It is therefore easier to base the construction of the marginal likelihood on the joint marginal survival function which can easily be derived making use of the simple Laplace transform (1.12). For the positive stable distribution expression (1.14) becomes

$$S(\boldsymbol{t}_{n_i}) = \exp\left[-\left(\sum_{j=1}^{n_i} H(t_{ij})\right)^{\theta}\right].$$

From the joint marginal survival function we can construct the marginal loglikelihood. For simplicity, we will only give the explicit expression for the marginal loglikelihood for bivariate data, needed in Section 3.4

$$\begin{aligned} l_{marg}(\boldsymbol{\zeta}) \;=\; & \sum_{i=1}^{k}\left[\sum_{j=1}^{2}\delta_{ij}\log\left(h_0\left(y_{ij}\right)\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right) + d_i\log\theta\right. \\ & -\left(\sum_{j=1}^{2}H_0(y_{ij})\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right)^{\theta} \\ & \left.+d_i(\theta-1)\log\left(\sum_{j=1}^{2}H_0(y_{ij})\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right) + C_{i,d_i}\right], \end{aligned}$$

with $d_i = \sum_{j=1}^{n_i}\delta_{ij}$, $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard. C_{i,d_i} is a term that depends on the number of events in the cluster. In case of zero events or one event $C_{i,0} = C_{i,1} = 0$, in case of two events $C_{i,2}$

takes the following form

$$C_{i,2} = \log\left(1 + \theta^{-1}(1-\theta)\left(\sum_{j=1}^{2} H_0(y_{ij})\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right)^{-\theta}\right).$$

If a parametric assumption is made for the baseline hazard, the marginal likelihood is fully parametric and classical maximum likelihood techniques can be used to estimate the parameters. Standard errors can be obtained from the inverse of the observed information matrix.
Estimation methods for the semiparametric positive stable frailty model can be found in Wang et al. (1995), Fine et al. (2003) and Martinussen and Pipper (2005).

1.8 The copula model

Copulas are functions that join or couple multivariate distribution functions to their onedimensional marginal distribution functions (Nelsen, 2006). Let X and Y be two random variables with $F(x) = P(X \leqslant x)$ and $G(y) = P(Y \leqslant y)$ the distribution functions of X and Y, respectively. Let $H(x, y) = P(X \leqslant x, Y \leqslant y)$ the joint distribution function. According to the theorem of Sklar (Sklar, 1959) there exists a copula C such that for all (x, y) $H(x, y) = C(F(x), G(y))$.
In this thesis we will focus on survival copulas. In a survival copula the copula function couples the marginal survival functions of the different subjects in a cluster and their joint survival function. Copula models are used to model clustered data with small and equal cluster size. An ordering in the cluster such that the first marginal survival function always refers to the same subject within a cluster (for example, the first marginal survival function is always the marginal survival function of the front left udder quarter) is also needed. Let n be the number of cluster members. Especially for bivariate data ($n = 2$) the copula approach is often used, although multivariate extensions are discussed in the literature (Massonnet et al., 2009).
The joint survival function is given by

$$S\left(\boldsymbol{t}_n\right) = S\left(t_1, \ldots, t_n\right) = C_{\boldsymbol{\theta}}\left(S_1(t_1), \ldots, S_n(t_n)\right)$$

with $\boldsymbol{t}_n = (t_1, \ldots, t_n)$ and $S_j(t), j = 1, \ldots, n$ the marginal survival functions of the n subjects in a cluster. These marginal survival functions are obtained from the marginal approach not taking into account the clustering in the data (see Section 4.3.1). $C_{\boldsymbol{\theta}}$ is the copula function defined on

$(v_1, \ldots, v_n) \in [0,1]^n$ and takes values in $[0,1]$. Its existence (and uniqueness if the marginal survival functions are all continuous) follows from Sklar's theorem (Sklar, 1959). One class of copulas are Archimedean copulas which will be used in Chapter 3. For an in-depth discussion on copulas see Nelsen (2006).
Inferential procedures for copula models typically use a two-stage approach (e.g., see Shih and Louis (1995b)). In the first stage the marginal survival functions are estimated (parametric, semiparametric or nonparametric estimation has been considered). In the second stage, estimates for the parameters in the copula function are obtained by maximization of the likelihood with respect to the copula function parameter, after we have replaced the marginal survival functions by the corresponding estimated versions (obtained in the first stage) in the likelihood expression. This procedure is based on the attractive feature of copula models that marginal distributions do not depend on the choice of the correlation structure, therefore, the marginal distributions and the correlation can be modeled separately.
When modeling the marginal survival functions in a semiparametric or nonparametric way the two-stage approach is a natural way to obtain parameter estimates. For marginal survival functions modeled in a parametric way maximum likelihood estimation for all the parameters simultaneously (i.e., the parameters of the marginal survival functions and the parameters of the copula) is also possible.

1.9 Data sets

In the following sections we present two data sets which will be used in the further chapters to demonstrate the developed methodology. The first example is a data set on the time to diagnosis of fracture healing in dogs by two different imaging techniques. The cluster is the dog and the two observation times (one for each imaging technique) are correlated within the dog. The second example is a data set on the time to infection in an udder quarter of a dairy cow. In this case the cow is the cluster and the individual observations at the udder quarter level are correlated within the cow. More details on the data sets can be found in the following sections.

1.9.1 Diagnosis data set

Medical imaging has become an important tool in the veterinary hospital to assess whether and when a fracture has healed. In dogs, the standard technique to evaluate fracture healing is based on radiography (RX). In

humans, however, fracture healing has been identified at an earlier stage by ultrasonography (US) than by RX. Furthermore, techniques based on US are cheaper and there would be no roentgen exposure of dogs and staff. To investigate the ability of US to diagnose fracture healing and compare the timing to diagnosis of fracture healing with RX in dogs, Risselada et al. (2005) set up a trial in which fracture healing is evaluated by both US and RX. In total, 106 dogs, treated in the veterinary university hospital of Ghent, are included in the trial and evaluated for time to diagnosis of fracture healing with the two techniques. Mean time to US identification of healing was 51 $\pm$ 23 days, range 5-107 days, for RX the mean time was 60 $\pm$ 27 days, range 1-163 days. Only 7 dogs are censored for time to diagnosis of fracture healing evaluated by RX; no censoring occurs for time to diagnosis of fracture healing evaluated by US. The censoring is due to the fact that dog owners do not show up anymore. The data for a few dogs are given in Table 1.1.

Table 1.1: Diagnosis data set. The first column contains the dog identification number, the second column gives the time (in days) to diagnosis, the third column gives the censoring status taking value one (status=1) if healing is observed and zero (status=0) otherwise. The last column gives the diagnostic technique (RX=radiography, US= ultrasonography).

Dogid	Time to diagnosis	Status	Method
1	63	1	RX
1	30	1	US
2	83	1	RX
2	83	1	US
...			
106	35	0	RX
106	35	1	US

1.9.2 Mastitis data set

An introduction to mastitis

Intramammary infection and mastitis

Mastitis, as a reaction to an intramammary infection (IMI), is economically the most important disease in the dairy sector of the western world because it is closely associated with reduced milk yield and milk quality. Control costs such as expenditures for treatment and preventive measures and extra labour time to execute treatment and preventive measures increase the total economic impact (Seegers et al., 2003). Therefore, control of IMI is an important component of dairy herd health programs. But, how should we approach and define IMI?
Mastitis, or inflammation of the mammary gland, is an innate defence mechanism (Burvenich et al., 2007) that is activated at the occasion of an IMI or udder injury. In the lactating cow mostly bacterial infections are concerned. Following invasion of bacteria through the teat canal and cistern, there is an abrupt influx of phagocytic cells from the circulation into the udder and its cisterns. These cells engulf and kill the invading organism. Invasion and phagocytosis is accompanied by the release of bioactive molecules that may be harmful for the udder and the cow. The clinical signs of mastitis are an expression of the host defence intended to destroy the invader and to repair the mammary tissue (Jain, 1979). There is, however, large variation in the clinical symptoms.
Cows can be diagnosed with mastitis on basis of the variation in these clinical symptoms.
Clinical mastitis is accompanied by swelling and pain of the udder, and a strong decline in milk production with changes in milk composition. Next to this, several degrees of general illness can be seen, such as, fever, general depression and decreased food intake. Clinical mastitis in cows can be easily diagnosed. In some cases, mastitis can occur without any apparent clinical symptom (subclinical mastitis). The diagnosis of subclinical mastitis has to be made in a laboratory on basis of the counting of the number of somatic cells (somatic cell count, SCC). The isolation of a pathogen can also be performed in the laboratory.
Theoretically, the internal environment of a healthy udder quarter is expected to be sterile; and consequently SCC should be equal to zero. However, phagocytes are always attracted from the circulation during milking because of minimal injury of the udder.

Under practical conditions, SCC in milk samples from healthy quarters is very low. However, there is no general agreement on the absolute SCC in these healthy milk samples; there is no precise computation. It is expected that the subjective probability for mastitis to occur is extremely low when SCC in a milk quarter sample is less than 100 000 cells per ml. The probability is close to 0. An elevation of SCC $>$ 100 000 cells per ml is an indication of mastitis. Whether accompanied by clinical signs or not, mastitis is always associated with an increase in SCC. Therefore, one tool to monitor udder health and milk quality is SCC (O'Brien et al., 1999).
Another tool is detecting the presence of a (specific) pathogen. This is based on the demonstration of an udder pathogen in a milk sample in the laboratory. Although logical reasoning expects that detection of such a pathogen in milk samples should represent the golden standard, it is not. For several reasons there are many false negative results. Due to specific properties of the pathogen and of the host, the pathogen can not always be isolated in the milk sample. For example, the cyclical shedding pattern of *Staphylococcus aureus* can cause a single milk sample to be negative while the udder quarter is infected. It is recommended to take and test two to three consecutive milk samples to increase the sensitivity (Sears et al., 1990). Milk samples of a suspected *Escherichia coli* (*E. coli*) infection are also often bacteriologically negative. The defence mechanism of the host is very effective against infection with *E. coli* and often the bacterium has already been removed at the time of sampling. In this thesis we focus on time to IMI with a specific pathogen.
The prevalence of an IMI is defined as the number of infected udder quarters at a given time divided by the total number of udder quarters (infected and uninfected) in the study. Incidence is a measure of the risk of an udder quarter to get infected within a specified period of time. It is defined as the number of newly infected udder quarters within a specified time period divided by the total number of udder quarters initially at risk. Thus, incidence conveys information about the risk of getting infected, whereas prevalence indicates how widespread the occurrence of infections is.

IMI and mastitis: an interaction of different factors

The detection of an IMI is a concurrence of three groups of actors : 1) the involved pathogen, 2) the management of the herd and 3) the physiology of the cow (see Figure 1.4).

The pathogens
Bacteria responsible for IMI's can be divided into two groups: contagious and environmental pathogens. Contagious pathogens are well adapted to

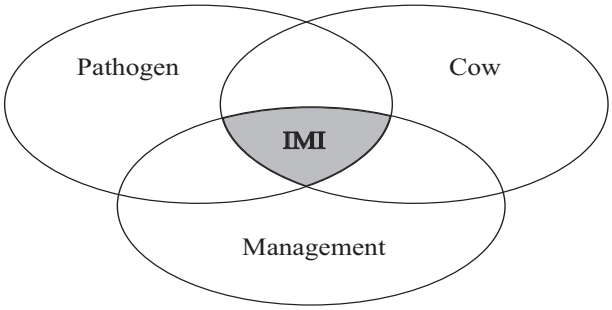

Figure 1.4: The interaction between the involved pathogen, the management of the herd and the physiology of the cow involved in the possibility to detect an intramammary infection

survival and colonization in the animal tissues. They frequently cause chronic subclinical IMI's. The main source of these pathogens in a dairy herd is the infected gland and transmission to uninfected udder quarters and cows occurs mainly during milking by contaminated milking equipment, milker's hands, cloths, etc... The primary source of environmental pathogens is the cow's surroundings. Environmental pathogens can not colonize the teat canal and need moisture and sticky dirt to breach the teat canal. This dichotomic classification model has been used during decades and has proved to be useful under practical conditions. However, over the last years some papers have defended the thesis that a gradual classification model would be more realistic. This is based on the fact that some environmental strains expressed some adaptation to the bovine udder.
The most important causal pathogens that induce contagious IMI's are *Staphylococcus aureus* (*Staph. aureus*) and *Streptococcus agalactiae* (*Strep. agalactiae*). Next to the fact that these IMI's tend to be chronic and subclinical, recurrent clinical episodes may occur.
IMI with *Strep. agalactiae* occurs mainly at the beginning and at the end of lactation. The most important sources of infection are the cow's udder and damaged teat skin. It spreads during milking through the equipment and is considered highly contagious. The bacterium has several virulence factors to resist defence mechanisms of the udder. Most infected cows show few clinical signs of mastitis but SCC is elevated and milk production de-

creased. The infection can also be acute with mild to moderate symptoms. *Strep. agalactiae* is very sensitive to penicillin and its survival outside the udder is limited. It has been eradicated from most herds in many countries. It will not be considered in this thesis.
Staph. aureus is more difficult to eradicate than *Strep. agalactiae*. A staphylococcol IMI can occur at all stages of lactation, but is most common during drying-off and immediately after calving. Staphylococci are able to actively colonize teat skin lesions and the teat canal. Once they adhered to the milk fat globules, the bacteria can move upwards in the udder. After entering the milk cisterns, staphylococci spread quickly in the udder. During chronic staphylococcal IMI, bugs may appear in epithelial cells, neutrophils and macrophages where they may be kept alive because of their ability to resist phagocytosis (Barrio et al., 2000). Therefore antibiotic therapy is often ineffective. Increased SCC and damaged milk secreting tissue occurs. The prognosis for *Staph. aureus* IMI is poor because only a few antibiotics exist that penetrate cells and affect the pathogens inside. Furthermore, their effect is poor. The more chronic the infection with *Sthaph. aureus* the poorer the prognosis. Peracute clinical mastitis, characterized by tissue oedema and necrosis can occur. In this thesis, infection with *Staph. aureus* will be considered.
Important environmental pathogens are *Escherichia coli* (*E. coli*) and *Streptococcus uberis* (*Strep. uberis*). The rate of IMI with *E. coli* is higher at the end of the dry period (colostrogenesis) and early lactation than during lactation. The udder quarter gets infected through faecal contamination in bedding material. Colonization of the teat canal is not a prerequisite and the bacteria are probably propelled directly through the teat canal. IMI with *E. coli* tends to be of short duration but can become chronic. IMI with *E. coli* is not further considered in this thesis.
IMI with *Strep. uberis* occurs usually at the beginning of lactation and at the end of the dry period. *Strep. uberis* is found in the cow's environment and on the cow's epithelia, faeces, teat skin, etc. Colonization of the teat canal is not a prerequisite and the bacteria are probably propelled directly through the teat canal. Some strains of *Strep. uberis* have capsules that alter phagocytosis. This would explain the poorer cure rate compared to the other Streptococci. The symptoms of *Strep. uberis* IMI are mild to moderate in most cases and SCC is elevated (Hoeben et al., 1999). IMI with *Strep. uberis* will be considered in this thesis.
Streptococcus dysgalactiae (*Strep. dysgalactiae*) is generally considered as environmental pathogen, but has also characteristics of a contagious pathogen. It doesn't fit in the classical dichotomic infection model. *Strep. dysgalactiae*

resides mainly in the cow's udder and in teat injuries, but also elsewhere in the cow and its surroundings. It is less contagious than *Strep. agalactiae* and adhesion in the udder epithelium is weaker. Infection with *Strep. dysgalactiae* is most apparent at the beginning of lactation. Clinical symtoms of *Strep. dysgalactiae* mastitis are more severe and the disease is often acute. IMI with *Strep. dysgalactiae* will be considered in this thesis.
Corynebacterium bovis (*C. bovis*) is considered as a minor pathogen. Increase in SCC is low and milk production losses are limited. It is rarely isolated in cases of clinical mastitis. IMI with *C. bovis* will also be considered in this thesis.

Management factors
Management factors influencing prevalence and incidence of IMI's are directed towards prevention of IMI occurrence and towards curing infected udder quarters or cows. A good manager needs to know the basics of the pathogenesis of IMI's as mentioned afore because management practices can be different according to the problem. Preventive measures against IMI with contagious pathogens are mostly related to the milking process: wearing gloves during milking, using automatic take-offs, using postmilking teat dipping, milking problem cows last and yearly inspection of the milking system (Dufour et al., 2011). Preventive measures against IMI with environmental pathogens concern the general hygiene and food management of the herd. They include providing dry bedding (preferably sand), cleaning the calving pen after each calving, keeping environmental temperatures cool (good ventilation), but also the use of techniques to keep cows standing following milking. Prevention of teat injuries is important for the prevention of IMI with contagious as well as environmental pathogens. The effectiveness of administrating antibiotics to treat IMI's depends on the pathogen. Dry cow therapy is recommended as a treatment for existing infections and as a preventive measure for new infections. Culling of chronically infected cows should be considered.

Cow factors
One could think that the bacterial characteristics alone are responsible to explain variation in IMI's. The cow is however not helpless against IMI's. Since an udder becomes infected through the teat canal, the teat canal is the first line of defence against IMI's. It represents a physical and chemical barrier against the penetration of pathogens, but when the teat opening is dilated the risk of infection may be higher. Epithelial desquamation and the shear stress of the milk flow are mechanisms that inhibit bacterial col-

onization. Neutrophils accumulate beneath and between the epithelium of the teat canal wall and fight off bacterial infections. Since teat end condition deteriorates with increasing parity (Neijenhuis et al., 2001), parity is considered an important cow-factor.
The cow's innate defence mechanisms consist of humoral and cellular components. Phagocytosis and intracellular killing by bovine neutrophils are important host defence mechanisms during mastitis caused by *Staph. aureus*. At our department the phagocytosis and overall killing of a non slime-producing *Staph. aureus* and its slime-producing variant was studied and it was concluded that the presence of slime was responsible for a decreased phagocytic ingestion and overall killing by these phagocytic cells (Barrio et al., 2000).
However, susceptibility of the mammary gland to new infections seems also to be affected by the physiology of the lactating cow. IMI's are markedly increased during early involution (drying off ; active involution, first week or two, highest incidence of new infections) and during the periparturient period (colostrogenesis) (Oliver and Sordillo, 1988). These periods coincide with unique local and systemic physiological phenomena that interrupt or induce lactation. Considerable changes in mammary tissue remodeling and nutritional demands occur that interfere with the defence system (eg. neutrophil function) of the udder. The increase in IMI's at drying off do not fall within the scope of this thesis.
Kehrli et al. (1989) hypothesized that an immunocompromised condition during the periparturient period, predisposes the dairy cow to new infections and/or the progression of subclinical mastitis into clinical disease. At that moment a cause and effect relationship between a faltering innate defence system and the development of intramammary infection was not proved. However, one year later it was also shown that the severity of clinical *E. coli* mastitis was correlated with the decreased production of reactive oxygen species (ROS) (Heyneman et al., 1990) and decreased chemotaxis (Lohuis et al., 1990) of circulating neutrophils, isolated before the intramammary infection.
It is unlikely that periparturient immunosuppression is the result of a single physiological factor; more likely, several entities act in concert, with profound effects on the function of many organ systems of the periparturient dairy cow. Their defence system is unable to modulate the complex network of innate immune responses, leading to incomplete resolution of the pathogen and the inflammatory reaction. During the last 30 years, most efforts have been focused on neutrophil chemotaxis, phagocytosis, and bacterial killing. How these functions modulate the clinical outcome of mastitis,

and how they can be influenced by hormones and metabolism has been the subject of intensive research (Burvenich et al., 2007).
It is now clear that the stage of lactation influences the function of the innate immune system and by this the susceptibility to disease (Burvenich et al., 2004, 2007). The impact of the physiology of the animal on periparturient inflammatory disease is most pronounced for IMI with *E.coli*, to a lesser extend with *Staph. aureus* (Dosogne et al., 2001) and of minimal importance with *S.uberis* (Hoeben et al., 1999).
The ageing cow is also more susceptible to IMI for the same reasons as afore-mentioned (Burvenich et al., 2003). Blood neutrophil function was higher in younger animals than in cows after their 4th parturition. The drop in neutrophil ROS production around parturition is more pronounced in multiparous cows (Mehrzad et al., 2002). The pronounced reduction in ROS production and viability in milk neutrophils of multiparous cows may be involved in the underlying mechanisms that make older animals more susceptible to periparturient infectious diseases. Moreover, white blood cell viability and oxidative burst have been found to be significantly different between primiparous cows and multiparous cows during the periparturient period.
In conclusion, it is clear that host factors have an effect on IMI's with several pathogens. Together with the characteristics of the pathogen and management factors, they explain variation in IMI's in the cow.

Studying IMI's and mastitis

Different types of research have been conducted to investigate incidence, prevalence and predicting factors of mastitis and/or IMI's. Case studies mainly describe individual cases. Experimental studies are conducted to investigate the causal effect of risk or protective factors (independent variables) on the outcome (the dependent variable). Epidemiological studies can not proof a causal effect between a risk or protective factor and the outcome. Epidemiological evidence can only show that this factor is associated (correlated) with a higher incidence of disease (mastitis or IMI) in the population exposed to that factor.

Case studies
Mastitis research may be based on individual case studies. In a case study the researcher reports, for example, the clinical picture, treatment and recovery process of a single case. In a mastitis study, for example, a veterinarian visits the farm and investigates the udder of a cow for clinical signs of mas-

titis and/or tries to isolate pathogens from a milk sample. The veterinarian can administer a certain treatment and the cow is followed up to see whether the treatment is successful. A case study is usually purely descriptive and concerns only the cow under study. No conclusions about the general population of dairy cows with mastitis or a (specific) IMI can be drawn. It can however lead to the formulation of an interesting research hypothesis based on the experiences with that cow.

Experimental studies
In an experimental study on mastitis/IMI (for an example of an experimental study see for instance Vangroenweghe et al. (2004)) a hypothesis (for example a hypothesis deducted in a case study) is tested. Experimental studies are always hypothesis driven. The hypothesis should be clearly identified before carrying out the experiment. The hypothesis consists of two parts: the null hypothesis states what the researcher is trying to reject, the alternative hypothesis formulates what the researcher is trying to proof. Preferably only one, or at least a very small number of hypotheses is formulated. For example, the experimental study of Vangroenweghe et al. (2004) hypothesized that the application of 2 different inoculum doses of *E. coli* elicits differences in the innate immune response (alternative hypothesis), the null hypothesis being no difference. An experimental study is carefully planned. Due to ethical and financial reasons the number of study subjects (the sample size) is often small. The needed number of study subjects is calculated based on information on the expected difference, the expected variability in the observations and the desired power. The study subjects are very comparable to reduce variability in the observations. For example, for cows, restrictions concerning breed, parity, diet, health status, SCC or milk production can be applied. Study subjects are randomized to a specific group (for example to a specific inoculum dose). Randomization is very important in an experimental study (Vangroenweghe et al., 2004).
An experimental study on mastitis/IMI has several advantages. Since the variability between the cows is low and group allocation is random, observed differences in outcome can be attributed to the experimentally induced difference between the groups. Confounding or missing important covariates are less encountered in experimental studies. In an experimental study the timing of, for example, inoculation is known exactly and the reaction to this inoculation can be monitored continuously or at least in small time intervals (e.g. daily). Therefore, there is less uncertainty in the observations. Definitions of, for example, mastitis or IMI are often clearly stated. A disadvantage is the limited (in number and in diversity) study population.

Extrapolation of the obtained results to the general population of dairy cows is usually not possible because the results depend on, for example, the considered breed and/or parity status and on the region or the season in which the experiment took place.

Epidemiological studies
In an epidemiological study prevalence and/or incidence of mastitis/IMI's and correlations between risk or protective factors and incidence of mastitis/IMI's are investigated. Usually large numbers of animals are included in the study. This has the advantage that the population is very heterogeneous and makes it possible to extrapolate the obtained results to the general population of dairy cows. There are also disadvantages. To obtain such large numbers of animals usually different farms are included in the study. The animals within a farm have some features in common, such as housing, diet, hygiene,.... Since IMI's are assessed at the udder quarter level and udder quarters are obviously clustered within cow, two hierarchical levels (udder quarters clustered within cow, cows clustered within herd) are present in the data. This needs to be addressed in the statistical analysis to obtain valid results. The heterogeneous population makes it also impossible to claim a causal effect between a risk or protective factor and the outcome. The considered factor could actually influence another (maybe unobserved) factor, which actually has a causal effect on the outcome. This is called confounding and is often a problem in epidemiological studies.
Usually also a large number of risk or protective factors are recorded at the udder quarter level, the cow level and the herd level. On the one hand this yields a lot of possibly useful information, on the other hand it is impossible to record everything and maybe useful factors are missing.
Many epidemiological studies on IMI's are cross-sectional studies (Dufour et al., 2011). The study population is examined only once. The researcher visits the farm(s) at a certain point in time, establishes the infection status of the udder quarters of the cows and collects data on, for example, parity, stage of lactation, diet, housing,... A cross-sectional study is quick, but the infection status of the udder quarters is only established at one point in time. There is no information on when exactly the infection happened and there is no information on the future infection status. Longitudinal studies are found less in the literature (for an example of a longitudinal study see for instance Lam et al. (1997)). A longitudinal study starts at a certain point in time and follows the study population until an infection occurs and often even thereafter, until the end of the study. This way a lot more information is available for the researcher. The timing of the infection can be

recorded (early in lactation, late in lactation), the evolution of the infection can be followed and risk factors that change over time can be recorded. On the down side, longitudinal studies are very time and money consuming and therefore less available. Furthermore it requires regular visits of the cows by the researcher. It is practically impossible to monitor all the cows continuously, therefore the exact time of infection is never known. It is only known that an udder quarter got infected between the last visit at which it was infection free and the first visit at which it was infected. This problem is often encountered in epidemiological research and has an important impact on statistical analysis conducted afterwards. To address this problem (called interval censoring) together with the problem of clustering in the data (hierarchical data) available statistical techniques have to be extended. The development of techniques that can handle clustering and interval censoring simultaneously is the main topic of this thesis. The developed methodology will be illustrated using an epidemiological study, described in the next section.

Mathematical modeling
To draw valid conclusions from conducted research, experimental or epidemiological, a proper statistical analysis of the observed data is essential. As Box (1976) eloquently put it "All models are wrong, but some are useful". Every model makes assumptions which often can not be verified and while no model is perfect for the data, an appropriate model can provide important insights in the studied matter. The statistical model should exploit the information in the data to its full extent and should model the specific data structure correctly. For example, the dynamic information in a longitudinal study should be fully exploited and not reduced to a static problem. Clustering in the data should be accounted for, otherwise obtained results are not valid. Often new statistical techniques need to be developed to meet specific data characteristics or specific research questions. Without attempting to create a perfect model (which does not exist) the developed methodology in this thesis addresses two characteristics often encountered in epidemiological studies: interval censoring caused by the dynamic character of the study and the clustering of udder quarters within cow.

The mastitis data

In an extensive study, 1207 cows are selected from twenty-five dairy herds located in the provinces of East and West Flanders, Belgium. They are followed up for infection at the udder quarter level during a 20-month period (February 1993 to September 1994) (Laevens et al., 1997).
Criteria for herd selection were willingness of the farmer to cooperate, participation in the DHI program that was organized by the Flemish Cattle Breeding Association, minimum herd size of 25 cows, and breed (Holstein Friesian and Red and White). The infection status of the udder quarter is determined on basis of isolation of a pathogen in a simple milk sample in the laboratory. Milk samples were taken by trained technicians. The teat ends were cleaned with dry udder cloths. Dirty teats were washed and dried. Before milk samples were taken, teats were disinfected with cotton moistened with a solution of ethyl alcohol (70%) and chlorhexidine (200mg/100ml). The milk samples were transported immediately after collection to a laboratory and streaked for initial isolation within two to three hours after collection. Quarter milk samples were streaked onto a 90-mm Petri dish with a blood agar base (Oxoid, Basingstoke, England) supplemented with 5% bovine blood; samples were also streaked onto an Edwards medium (Oxoid) supplemented with 5% bovine blood. Agar plates were incubated at 37°C and read after 24h and 48h. Isolates were classified as described by the National Mastitis Council (Harmon et al., 1990). The data set contains information on times to infection with different bacteria: *Staph. aureus*, *Strep. dysgalactiae*, *Strep. agalactiae*, *Strep. uberis*, coliforms (*E. coli*, *Klebsiella*), *C. bovis*, ... Four bacteria are selected (*Staph. aureus*, *C. bovis*, *Strep. dysgalactiae* and *Strep. uberis*) and considered in this thesis. Data on times to infection with these bacteria are used to illustrate the developed methodology.
Since the data set contains information on times to infection with a bacterium survival analysis techniques are the appropriate tool to model them. Udder quarter infection data have often been reduced to binary data (Zadoks et al., 2001): either an infection occurs in the udder quarter or not during the complete lactation period. However, this reduces the amount of information considerably: all information on the timing of the infection is lost.
Cows are monthly screened for bacterial infections at the udder quarter level from the time of parturition, at which the cow was included in the study and assumed to be infection free, until the end of the lactation period or the end of the study period. However, due to a lack of staff in summer months, cows are screened only in July or August meaning that at least one

interval spans two months. Furthermore, the time between two successive visits is not always exactly one month and the recorded visiting times for each cow differ depending on when they were included in the study. Due to the periodic follow-up, observations for udder quarters that experience an event are interval-censored with lower bound the last visit with a negatively tested milk sample and upper bound the first visit with a positively tested milk sample. Observations can be right-censored if no infection has occurred before the end of the lactation period, which is roughly 300-350 days but different for every cow, before the end of the study or if the cow is lost to follow-up during the study, for example due to culling. Each udder quarter is separated from the three other quarters. Therefore, one quarter might be infected while the other quarters remain infection-free. However, since udder quarters are clustered within cow, observations from different udder quarters of one cow can not assumed to be independent. Cows are further clustered within herd.

The interval censoring and clustering in the data are two characteristics of the mastitis data that need to be addressed in the statistical analysis. Survival analysis techniques for univariate (unclustered) data are available in the literature and commercial software packages (Sun, 2006). Also, the use of appropriate techniques for clustered time to event data is widespread (Wei and Glidden, 1997; Hougaard, 1999; Kelly and Lim, 2000). However, techniques that address the interval censoring and clustering in data simultaneously are found less in the literature. The aim of this thesis is to develop new techniques to model data that are simultaneously clustered and interval-censored, two important characteristics present in the mastitis data. The udder quarter is considered as the observational unit, clustered within cow. Different covariates are recorded in the mastitis study: SCC, bedding, parity, properties of the milking equipment, duration of milking, temperature, herdsize, etc. The aim of this thesis was to develop new methodology for interval-censored, clustered data; not to give a full risk factor analysis of the mastitis data. We therefore select two relevant covariates for illustrative purposes. Two types of covariates are considered.

Cow level covariates take the same value for every udder quarter of the cow (e.g. number of calvings or parity). Several studies have shown that prevalence as well as incidence of intra mammary infections increases with parity (Vecht et al., 1989; Weller et al., 1992). Several hypotheses have been suggested to explain these findings, e.g. teat end condition deteriorates with increasing parity (Neijenhuis et al., 2001). Because the teat end is a physical barrier that prevents organisms from invading the udder, impaired teat ends make the udder more vulnerable for intra mammary infections. Following

Table 1.2: Mastitis data set. The first column contains the cow identification number, the second, third, and fourth columns contain the time (in days) to infection with *Staphylococcus aureus* (the lower bound, upper bound, and midpoint of the interval, resp.), the fifth column gives the censoring status taking value one (status=1) if infection is observed and zero (status=0) otherwise. The last two columns give the parity (primiparous cow (parity=0) or multiparous cow (parity=1)) and the udder quarter (RL=Rear-Left, FL=Front-Left, RR=Rear-Right, FR=Front-Right).

	Time to infection					
Cowid	Lower	Upper	Midpoint	Status	Parity	Quarter
1	84	154	119	1	0	RL
1	50	84	67	1	0	FL
1	50	84	67	1	0	RR
1	50	84	67	1	0	FR
2	134	160	147	1	1	RL
2	44	106	75	1	1	FL
2	134	160	147	0	1	RR
2	134	160	147	0	1	FR
...						
1206	221	248	234.5	0	1	RL
1206	221	248	234.5	0	1	FL
1206	221	248	234.5	0	1	RR
1206	221	248	234.5	0	1	FR
1207	247	279	263	0	0	RL
1207	247	279	263	0	0	FL
1207	247	279	263	1	0	RR
1207	247	279	263	0	0	FR

categories for the parity covariate are available: (i) primiparous cows (one calving, parity = 0), (ii) cows with between two and four calvings (parity = 1) and (iii) cows with more than four calvings (parity = 2). In some examples parity will be dichotomized into primiparous cows (parity=0) and multiparous cows (parity=1) for reasons of simplicity.
Udder quarter level covariates change within the cow (e.g. position of the udder quarter, front or rear). The difference in teat end condition between front and rear quarters has also been put forward to explain the difference in infection status (Adkinson et al., 1993; Barkema et al., 1997; Schepers et al., 1997). A subset of the data for infection with *Staph. aureus* is presented in Table 1.2.

Chapter 2

Research objectives

The research conducted in this thesis was inspired by the mastitis data set given in Section 1.9.2. Specific characteristics of the mastitis data and specific research questions for the mastitis data required the development of new statistical techniques to analyze the data. The specific objectives of this thesis were

- To evaluate differences and similarities between existing methodologies to model clustered survival data
- To develop new techniques to model fourdimensional clustered, interval-censored data based on the shared frailty model
- To develop new techniques to model different correlation structures between the udder quarters based on the correlated frailty model
- To apply existing and newly developed statistical tools to gain insight in the mastitis data, especially where it concerns the clustering aspect

Comparing existing methodology includes a comparison of the shared frailty model and the copula model for right-censored data, revealing several differences and similarities between the two models. This comparison is described in Chapter 3. The comparison also includes a review of the most frequently used methods to analyze clustered, interval-censored data: the marginal model, the fixed effects model and the copula model are considered and we point out some disadvantages or shortcomings of these models in particular for the mastitis data. The implementation and use of these models in commercial software packages is also discussed. This review can be found in Chapter 4.

To model fourdimensional clustered, interval-censored data we propose an extension of three models for right-censored data to interval-censored data: the shared gamma frailty model, a specific copula model and the correlated gamma frailty model. This approach assumes a correlation structure, such that the correlation between any pair of two event times in the four udder quarters is the same (symmetric correlation structure). Chapter 5 presents the extension of the shared gamma frailty model, the copula model for interval-censored data is described in Chapter 4 and the correlated gamma frailty model for interval-censored data is discussed in Chapter 6.

To investigate the correlation structure between the udder quarters further and to allow more complex correlation structures than the symmetric one, fourdimensional correlated gamma frailty models are proposed that allow

different correlations between the udder quarters. These models are described in Chapter 6.

Application of the different models to the mastitis data provides answers to questions concerning the effect of different types of covariates (at the cow and udder quarter level). The newly developed models also provide information on the most likely correlation structure, and on the size of the correlation between the event times, clustered in the cow udder. Results of the analysis of the mastitis data can be found in the chapters presenting the used methodology.

Chapter 3

Similarities and differences between the shared frailty and copula model

Based on:
Goethals, K., Janssen, P., and Duchateau, L. (2008), "Frailty models and copulas: similarities and differences," *Journal of Applied Statistics*, 35, 1071-1079.
Goethals, K., Janssen, P., and Duchateau, L. (2011), "Frailties and copulas, not two of a kind," *Risk and Decision Analysis*, accepted for publication.

3.1 Introduction

It is often claimed in the literature (Manatunga and Oakes, 1999; Viswanathan and Manatunga, 2001; Andersen, 2005) that there is equivalence between an Archimedean copula model and a shared frailty model with a particular frailty density, but in this chapter we will demonstrate that this claim is in most cases incorrect. For the theoretical discussion we will restrict to bivariate survival data, i.e., clustered survival data with clusters of size two, to keep notation simple. To illustrate our point the diagnosis data set of Section 1.9.1, which contains bivariate data, will be used. Throughout the theoretical discussion we will refer to this data set to make things more clear. Our findings will also be confirmed using the *Corynebacterium bovis* infection data set, which consists of 1196 clusters of four observations. Therefore, the copula likelihood for bivariate data needs to be extended to handle clusters of four observations. To deal with the interval censoring in the data imputation of the midpoint will be used.

3.2 The copula and the frailty model

Consider the two clustered diagnostic times (T_1, T_2) (T_1 for RX, T_2 for US) and let $S_{1,c}(t)$ and $S_{2,c}(t)$ be the marginal survival functions for the RX and US imaging technique. The subindex c is added to denote that the joint survival function is obtained from the copula presentation. For a twodimensional survival copula model the joint survival function is given by

$$S_c(t_1, t_2) = C_\theta \{S_{1,c}(t_1), S_{2,c}(t_2)\},$$

with C_θ a copula function, i.e., a function on the unit square
$C_\theta : [0,1]^2 \rightarrow [0,1] : (u,v) \rightarrow C_\theta(u,v)$ parameterized by θ (possibly a vector).
The shared frailty model, on the other hand, is given by

$$h_{ij}(t) = z_i h_{j,z}(t),$$

with $h_{ij}(t)$ the hazard at time t in cluster i, $i = 1 \ldots, k$, for diagnosis technique j (1=RX, 2=US), $h_{j,z}(t)$ the conditional hazard at time t for a cluster with frailty equal to one and diagnosis technique j and z_i the frailty term.
To compare copula models and shared frailty models we consider the family of Archimedean copulas

$$C_\theta(u,v) = p\{q(u) + q(v)\},$$

where $p(.)$ is any nonnegative decreasing function with $p(0) = 1$ and nonnegative second derivative and $q(.)$ is its inverse function; $p(.)$ is parameterized by θ. To make the link between copula and shared frailty models, we consider functions $p(.)$ that are Laplace transforms of frailty densities $f_Z(.)$

$$p(s) = \mathcal{L}(s) = \mathrm{E}\left\{\exp\left(-Zs\right)\right\} = \int_0^\infty \exp\left(-zs\right) f_Z(z)dz$$

leading to

$$C_\theta(u, v) = \mathcal{L}\left\{\mathcal{L}^{-1}(u) + \mathcal{L}^{-1}(v)\right\}.$$

For the copula model the joint survival function is

$$S_c\left(t_1, t_2\right) = \mathcal{L}\left[\mathcal{L}^{-1}\left\{S_{1,c}(t_1)\right\} + \mathcal{L}^{-1}\left\{S_{2,c}(t_2)\right\}\right]. \tag{3.1}$$

For the shared frailty model the joint conditional survival function for cluster i is $S_i\left(t_1, t_2\right) = \exp\left[-z_i\left\{H_{1,z}\left(t_1\right) + H_{2,z}\left(t_2\right)\right\}\right]$ with $H_{j,z}(t) = \int_0^t h_{j,z}(s)ds$. The joint survival function can be obtained by integrating out the frailties with respect to the frailty density

$$\begin{aligned} S_f\left(t_1, t_2\right) &= \int_0^\infty S_i\left(t_1, t_2\right) f_Z(z_i)dz_i \\ &= \mathrm{E}\left[\exp\left\{-Z\left(H_{1,z}(t_1) + H_{2,z}(t_2)\right)\right\}\right]. \end{aligned} \tag{3.2}$$

The subindex f is added to denote that the joint survival function is obtained from the conditional frailty model.
The joint survival function derived from the frailty model (3.2) and the joint survival function specified for the copula model (3.1) are two different ways to model $\mathrm{P}\left(T_1 > t_1, T_2 > t_2\right)$.
Expression (3.2) is nothing but the Laplace transform of the frailty distribution evaluated at $s = H_{1,z}\left(t_1\right) + H_{2,z}\left(t_2\right)$ so that

$$S_f\left(t_1, t_2\right) = \mathcal{L}\left\{H_{1,z}\left(t_1\right) + H_{2,z}\left(t_2\right)\right\}. \tag{3.3}$$

Furthermore, the marginal survival function for each of the two imaging techniques can be obtained by putting the diagnosis time for the other diagnostic technique equal to zero in (3.3) and thus $S_{j,f}(t) = \mathcal{L}\left\{H_{j,z}\left(t\right)\right\}$. It follows that

$$H_{j,z}(t) = \mathcal{L}^{-1}\left\{S_{j,f}(t)\right\}. \tag{3.4}$$

Using this relationship, (3.3) can be written as

$$S_f\left(t_1, t_2\right) = \mathcal{L}\left[\mathcal{L}^{-1}\left\{S_{1,f}\left(t_1\right)\right\} + \mathcal{L}^{-1}\left\{S_{2,f}\left(t_2\right)\right\}\right]. \tag{3.5}$$

Remark that the correlation structure used to obtain the joint survival function from the marginal survival functions in expressions (3.1) and (3.5) is the same. The arguments of the correlation structure, the marginal survival functions, however, are not the same. From (3.1) and (3.5) it follows that the two models are different in nature. This will be demonstrated in the next section, where we compare the Clayton-Oakes copula with Weibull marginal survival functions as arguments and the shared gamma frailty model with conditional Weibull hazards. In Section 3.4 a similar comparison for the positive stable copula and the shared frailty model with positive stable frailty density shows the exceptional character of this model, in the sense that both models are the same.

3.3 The Clayton-Oakes copula and the gamma frailty model

3.3.1 The diagnosis data

Assume that the marginal survival functions in the copula model are obtained from Weibull hazards and use the two-stage approach of Shih and Louis (1995b) to obtain parameter estimates. In the first step, parameter estimates for λ_j and γ_j are obtained by fitting the following survival model in each group (RX or US) separately

$$h_{j,c}(t) = \lambda_j \gamma_j t^{\gamma_j - 1}, \tag{3.6}$$

with $j = 1$ for the RX diagnosis and $j = 2$ for the US diagnosis. The parameter estimates (ML estimates) for the diagnosis data set are $\hat{\lambda}_1 = 0.106$ (0.024), $\hat{\gamma}_1 = 2.539(0.191)$, $\hat{\lambda}_2 = 0.219(0.039)$ and $\hat{\gamma}_2 = 2.323(0.175)$. To model the correlation we use the joint survival function (3.1) with

$$\mathcal{L}(s) = (1 + \theta s)^{-1/\theta} \text{ and } \mathcal{L}^{-1}(s) = (s^{-\theta} - 1)/\theta \tag{3.7}$$

with $\theta \geqslant 0$. $\mathcal{L}(.)$ is the Laplace transform of the one-parameter gamma distribution (1.9). The corresponding copula $C_\theta(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$ is the Clayton-Oakes copula (Clayton, 1978; Oakes, 1982).
The joint survival function then becomes

$$S_c\left(t_1, t_2\right) = \left[\left\{S_{1,c}(t_1)\right\}^{-\theta} + \left\{S_{2,c}(t_2)\right\}^{-\theta} - 1\right]^{-1/\theta}. \tag{3.8}$$

Based on the joint survival function (3.8) the likelihood can be constructed (see e.g. Shih and Louis (1995b)):

$$\begin{aligned} L(\boldsymbol{\zeta}) &= \prod_{i=1}^{k} \left(f_c(y_{i1}, y_{i2})\right)^{\delta_{i1}\delta_{i2}} \left(-\frac{\partial S_c(y_{i1}, y_{i2})}{\partial y_{i1}}\right)^{\delta_{i1}(1-\delta_{i2})} \\ &\quad \left(-\frac{\partial S_c(y_{i1}, y_{i2})}{\partial y_{i2}}\right)^{(1-\delta_{i1})\delta_{i2}} \left(S_c(y_{i1}, y_{i2})\right)^{(1-\delta_{i1})(1-\delta_{i2})}, \end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard. So we have four different possible contributions, depending on the censoring status of the two subjects in the cluster. A cluster with two censored subjects has contribution $L_{i,(0,0)} = S_c(y_{i1}, y_{i2})$, a cluster with two event times has contribution $L_{i,(1,1)} = f_c(y_{i1}, y_{i2})$; the contribution of a cluster with one event time and one censored observation is $L_{i,(1,0)} = -\frac{\partial S_c(y_{i1}, y_{i2})}{\partial y_{i1}}$, respectively $L_{i,(0,1)} = -\frac{\partial S_c(y_{i1}, y_{i2})}{\partial y_{i2}}$, if we observe an event time for the first (second) subject and a censored observation for the second (first) subject.
The likelihood contributions of the different clusters for a Clayton copula are given by

$$L_{i,(0,0)} = \left[\{S_{1,c}(y_{i1})\}^{-\theta} + \{S_{2,c}(y_{i2})\}^{-\theta} - 1\right]^{-1/\theta},$$

for clusters with two censored observations,

$$L_{i,(1,0)} = \left[\{S_{1,c}(y_{i1})\}^{-\theta} + \{S_{2,c}(y_{i2})\}^{-\theta} - 1\right]^{-1/\theta-1} \{S_{1,c}(y_{i1})\}^{-\theta-1} f_{1,c}(y_{i1}),$$

for clusters with an event for the first subject and a censored observation for the second subject,

$$L_{i,(0,1)} = \left[\{S_{1,c}(y_{i1})\}^{-\theta} + \{S_{2,c}(y_{i2})\}^{-\theta} - 1\right]^{-1/\theta-1} \{S_{2,c}(y_{i2})\}^{-\theta-1} f_{2,c}(y_{i2}),$$

for clusters with an event for the second subject and a censored observation for the first subject, and finally

$$\begin{aligned} L_{i,(1,1)} &= (1+\theta)\left[\{S_{1,c}(y_{i1})\}^{-\theta} + \{S_{2,c}(y_{i2})\}^{-\theta} - 1\right]^{-1/\theta-2} \\ &\quad \{S_{1,c}(y_{i1})\}^{-\theta-1} \{S_{2,c}(y_{i2})\}^{-\theta-1} f_{1,c}(y_{i1}) f_{2,c}(y_{i2}), \end{aligned}$$

for clusters with two events.
In the second step we replace in the likelihood $S_{j,c}(.)$ by $\hat{S}_{j,c}(.)$, obtained by

replacing λ_j, γ_j by $\hat{\lambda}_j$, $\hat{\gamma}_j$ (for $j = 1, 2$), and we then maximize the likelihood with respect to θ. In our example $\hat{\theta}$ is 0.890 (0.203). We will interpret the estimate for the parameter θ through its relationship with Kendall's τ. Kendall's τ is a global measure of correlation, defined as

$$\tau = \mathrm{P}\left((T_{i1} - T_{k1})(T_{i2} - T_{k2}) \rangle\, 0\right) - \mathrm{P}\left((T_{i1} - T_{k1})(T_{i2} - T_{k2}) < 0\right),$$

with (T_{i1}, T_{i2}), (T_{k1}, T_{k2}) the event times in two randomly chosen pairs. Values for τ are between -1 and 1, 1 corresponding to a perfect correlation, -1 meaning a perfect inverse correlation. If Kendall's τ is equal to 0, the event times are independent. The relationship between Kendall's τ and θ in the Clayton copula is given by $\tau = \theta/(\theta + 2)$. Therefore, in our example $\hat{\tau} = 0.308$. Since the marginal survival functions and the copula are modeled in a parametric way, the likelihood obtained from the joint survival function can also be maximized jointly for the marginal survival function parameters and the copula function parameter, leading to parameter estimates $\hat{\lambda}_1 = 0.145$ (0.030), $\hat{\gamma}_1 = 2.341$ (0.181), $\hat{\lambda}_2 = 0.233$ (0.042), $\hat{\gamma}_2 = 2.212$ (0.181) and $\hat{\theta} = 1.066$ (0.308) (see Table 3.1). The estimate of Kendall's τ is 0.348. Durrleman et al. (2000) give a detailed comparison between the two-stage approach and the (joint) maximization of the likelihood.
For the frailty model we start from a conditional Weibull hazard with different $\widetilde{\lambda}$ and $\widetilde{\gamma}$ parameters for the two diagnostic techniques (this is similar to the way in which the marginal survival functions in the copula model were modeled)

$$h_{ij}(t) = z_i \widetilde{\lambda}_j \widetilde{\gamma}_j t^{\widetilde{\gamma}_j - 1}, \tag{3.9}$$

with $z_1, \ldots, z_k$ independent realizations of the one parameter gamma distribution with mean one and variance $\widetilde{\theta}$ (see (1.9)).
The Laplace transform for the gamma distribution and its inverse is given in (3.7). Plugging (3.7) into (3.3) leads to the joint survival function

$$S_f\left(t_1, t_2\right) = \left[1 + \widetilde{\theta}\left\{H_{1,z}(t_1) + H_{2,z}(t_2)\right\}\right]^{-1/\widetilde{\theta}}.$$

Making use of (3.4) this can be rewritten as

$$\begin{aligned} S_f\left(t_1, t_2\right) &= \left[1 + \left[\left\{S_{1,f}(t_1)\right\}^{-\widetilde{\theta}} - 1\right] + \left[\left\{S_{2,f}(t_2)\right\}^{-\widetilde{\theta}} - 1\right]\right]^{-1/\widetilde{\theta}} \\ &= \left[\left\{S_{1,f}(t_1)\right\}^{-\widetilde{\theta}} + \left\{S_{2,f}(t_2)\right\}^{-\widetilde{\theta}} - 1\right]^{-1/\widetilde{\theta}}. \end{aligned}$$

This expression looks similar to the copula form representation in (3.8). There is, however, the substantial difference that $S_{j,f}(t) \neq S_{j,c}(t)$, $j =$

1, 2. The marginal survival function $S_{j,f}(t) = \left(1 + \widetilde{\theta}\widetilde{\lambda}_j t^{\widetilde{\gamma}_j}\right)^{-1/\widetilde{\theta}}$ is not of the Weibull form. Note that the frailty parameter also shows up in $S_{j,f}(.)$.
Parameter estimates for shared frailty models with a parametric baseline function can be easily obtained through maximization of the observable likelihood (see Section 1.7.5). Estimates for the parameters $\widetilde{\lambda}_1$, $\widetilde{\gamma}_1$, $\widetilde{\lambda}_2$, $\widetilde{\gamma}_2$ and $\widetilde{\theta}$ of the frailty model are given by 0.079 (0.021), 3.827 (0.369), 0.218 (0.046), 3.456 (0.333) and 0.909 (0.266), respectively (see Table 3.1). The estimate for Kendall's τ is 0.312.
The parameter estimates for $\widetilde{\lambda}_j$, $\widetilde{\gamma}_j$ are the parameter estimates from the conditional hazard whereas the parameter estimates from the copula model refer to the marginal hazard and survival functions. To compare the two models, note that the marginal hazard function in the frailty model is given by

$$h_{j,f}(t) = \widetilde{\lambda}_j \widetilde{\gamma}_j t^{\widetilde{\gamma}_j - 1} \left(1 + \widetilde{\theta}\widetilde{\lambda}_j t^{\widetilde{\gamma}_j}\right)^{-1}, \tag{3.10}$$

whereas the marginal hazard in the copula model is given by (3.6) and does not contain θ. Figure 3.1 shows the marginal hazards in the copula model and in the frailty model, the picture uses the estimated parameters. For $\widetilde{\gamma}_j > 1$, as is the case in our example, the conditional hazard in (3.9) is a monotone increasing function. With $\widetilde{\gamma}_j > 1$ the marginal hazard function (3.10) reaches a maximum in $t = \left\{(\widetilde{\gamma}_j - 1)/(\widetilde{\theta}\widetilde{\lambda}_j)\right\}^{1/\widetilde{\gamma}_j}$. The marginal hazard in the copula model is monotone increasing. Therefore, the marginal hazard functions in the two models can never be the same.
We also fitted the copula and frailty model using the semiparametric (Cox) model and the nonparametric model for the (conditional) hazards.
The semiparametric copula model is given by

$$h_{j,c}(t) = \begin{cases} h_0(t) & \text{for RX} \\ h_0(t)\exp(\beta) & \text{for US} \end{cases}$$

and the nonparametric copula model by

$$h_{j,c}(t) = \begin{cases} h_1(t) & \text{for RX} \\ h_2(t) & \text{for US} \end{cases}$$

with $h_0(t)$, $h_1(t)$ and $h_2(t)$ unspecified hazard functions. Estimation for the semiparametric and nonparametric copula model is typically based on the two-stage approach (Shih and Louis, 1995b; Spiekerman and Lin, 1998; Andersen, 2005). For the semiparametric model, we obtain in the first stage an estimate of β through partial likelihood maximization and we use

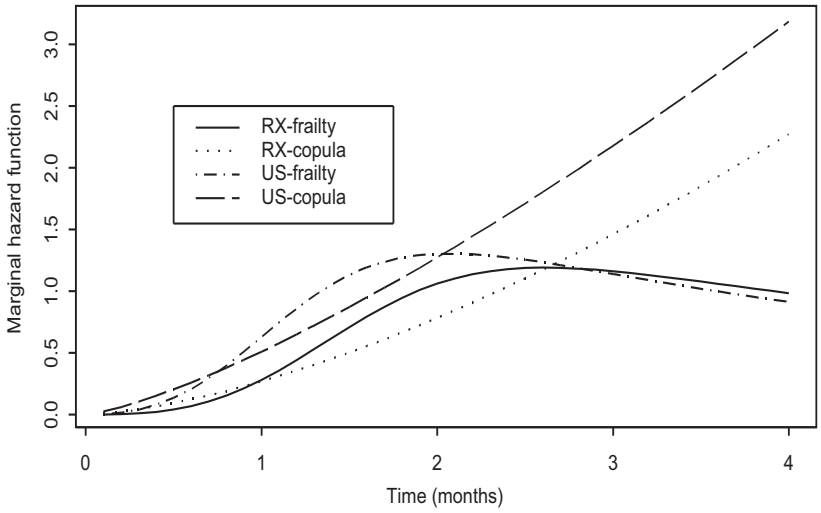

Figure 3.1: The marginal hazard functions from the shared frailty model and the copula model with gamma frailty density for the time to diagnosis of being healed data assessed by either US or RX.

the Breslow estimator of $S_0(t) = \exp(-\int_0^t h_0(t))$ (Breslow, 1974). For the nonparametric approach, we use the Nelson-Aalen estimator of $S_{j,c}(t) = \exp(-\int_0^t h_j(t))$, $j = 1, 2$ (Nelson, 1972; Aalen, 1978). In the second stage we replace the marginal survival functions in the likelihood by their corresponding estimates and then maximize with respect to θ.
The semiparametric frailty model is given by

$$h_{ij}(t) = \begin{cases} z_i \widetilde{h}_0(t) & \text{for RX} \\ z_i \widetilde{h}_0(t) \exp(\widetilde{\beta}) & \text{for US} \end{cases}$$

and the nonparametric frailty model by

$$h_{ij}(t) = \begin{cases} z_i \widetilde{h}_1(t) & \text{for RX} \\ z_i \widetilde{h}_2(t) & \text{for US} \end{cases}$$

with again $\widetilde{h}_0(t)$, $\widetilde{h}_1(t)$ and $\widetilde{h}_2(t)$ unspecified hazard functions.
Estimation for the semiparametric frailty model is based on the EM-algorithm (see Section 1.7.5). Estimation for the nonparametric frailty model is also based on the EM-algorithm but introducing imaging technique as stratification factor.
Parameter estimates in the semiparametric copula model are $\hat{\beta} = 0.508$ (0.087) and $\hat{\theta} = 0.997$ (0.193) with $\hat{\tau} = 0.333$; in the semiparametric (Cox) gamma frailty model estimates are given by $\hat{\widetilde{\beta}}$= 0.828 (0.164) and $\hat{\widetilde{\theta}} = 1.246$ (0.310) with $\hat{\tau} = 0.384$. In the nonparametric copula approach the estimate

Table 3.1: Diagnosis data: Estimates and their standard errors (Est (SE)) of the copula function parameter and the Weibull parameters of the Clayton copula model (using the two-stage approach or joint estimation) and of the shared gamma frailty model.

Parameter	Frailty model	Parameter	Copula model	
			two-stage	joint estimation
$\widetilde{\lambda}_{RX}$	0.079 (0.021)	λ_{RX}	0.106 (0.024)	0.145 (0.030)
$\widetilde{\gamma}_{RX}$	3.827 (0.369)	γ_{RX}	2.539 (0.191)	2.341 (0.181)
$\widetilde{\lambda}_{US}$	0.218 (0.046)	λ_{US}	0.219 (0.039)	0.233 (0.042)
$\widetilde{\gamma}_{US}$	3.456 (0.333)	γ_{US}	2.323 (0.175)	2.212 (0.181)
$\widetilde{\theta}$	0.909 (0.266)	θ	0.890 (0.203)	1.066 (0.308)

for θ is 1.236 (0.219), $\hat{\tau} = 0.382$; in the nonparametric gamma frailty model the estimate for $\widetilde{\theta}$ is 1.210 (0.289), $\hat{\tau} = 0.377$.
From the copula model as well as the frailty model it can be concluded that there is substantial positive correlation between the diagnosis times obtained by RX and US. Both models show that a fracture healing can be diagnosed earlier using US. The estimate of β in the copula model needs to be interpreted at the marginal level: diagnosis of fracture healing is earlier with US comparing two random dogs. The interpretation of the estimate of $\widetilde{\beta}$ in the frailty model is at the conditional level: diagnosis of fracture healing is earlier with US comparing two dogs with the same frailty. At the marginal level in the frailty model it can be seen that the hazard of diagnosis of fracture healing is smaller for RX than for US (meaning that the diagnosis of fracture healing will take longer using RX) until a certain point in time after which the hazard of diagnosis of fracture healing is slightly higher for RX. It is clear that the hazards are no longer proportional at the marginal

level.

3.3.2 The *Corynebacterium bovis* infection data

We will again assume Weibull hazards and use the two-stage approach of Shih and Louis (1995b) in the copula model to obtain parameter estimates. In the first step, parameter estimates for λ_j and γ_j are obtained by fitting the model (3.6) in each group separately. Now there are four groups corresponding to the front left udder quarters (FL, j=1), the rear left udder quarters (RL, j=2), the front right udder quarters (FR, j=3) and the rear right udder quarters (RR, j=4). The parameter estimates for the infection data set are $\hat{\lambda}_1 = 0.148$ (0.010), $\hat{\gamma}_1 = 1.305$ (0.052), $\hat{\lambda}_2 = 0.128$ (0.009), $\hat{\gamma}_2 = 1.310$ (0.055), $\hat{\lambda}_3 = 0.157$ (0.010), $\hat{\gamma}_3 = 1.255$ (0.049), $\hat{\lambda}_4 = 0.139$ (0.010) and $\hat{\gamma}_4 = 1.264$ (0.052) (see Table 3.2).
The fourdimensional joint survival function is

$$\begin{aligned} S_c\left(t_1,t_2,t_3,t_4\right) &= \Big[\{S_{1,c}(t_1)\}^{-\theta} + \{S_{2,c}(t_2)\}^{-\theta} \\ &\quad + \{S_{3,c}(t_3)\}^{-\theta} + \{S_{4,c}(t_4)\}^{-\theta} - 3\Big]^{-1/\theta}. \end{aligned} \tag{3.11}$$

Based on the joint survival function (3.11) the likelihood can be constructed (Massonnet et al., 2009). In the second step we again replace in the likelihood $S_{j,c}(.)$ by $\hat{S}_{j,c}(.)$, obtained by replacing λ_j, γ_j by $\hat{\lambda}_j$, $\hat{\gamma}_j$ (for $j = 1,2,3,4$), and we then maximize the likelihood with respect to θ. In this example $\hat{\theta}$ is 3.055 (0.124), $\hat{\tau} = 0.604$.
Joint maximization of the likelihood for the marginal survival function parameters and the copula function parameter leads to the following parameter estimates $\hat{\lambda}_1 = 0.141$ (0.009), $\hat{\gamma}_1 = 1.281$ (0.048), $\hat{\lambda}_2 = 0.121$ (0.008), $\hat{\gamma}_2 = 1.298$ (0.052), $\hat{\lambda}_3 = 0.150$ (0.010), $\hat{\gamma}_3 = 1.246$ (0.046), $\hat{\lambda}_4 = 0.128$ (0.009), $\hat{\gamma}_4 = 1.251$ (0.049), and $\hat{\theta} = 3.277$ (0.186) (see Table 3.2), $\hat{\tau} = 0.604$.
For the frailty model different $\widetilde{\lambda}$ and $\widetilde{\gamma}$ parameters are assumed for the four quarters in the conditional Weibull hazards. The joint survival function takes the form

$$\begin{aligned} S_f\left(t_1,t_2,t_3,t_4\right) &= \Big[\{S_{1,f}(t_1)\}^{-\widetilde{\theta}} + \{S_{2,f}(t_2)\}^{-\widetilde{\theta}} \\ &\quad + \{S_{3,f}(t_3)\}^{-\widetilde{\theta}} + \{S_{4,f}(t_4)\}^{-\widetilde{\theta}} - 3\Big]^{-1/\widetilde{\theta}}. \end{aligned}$$

Estimates for the parameters $\widetilde{\lambda}_1$, $\widetilde{\gamma}_1$, $\widetilde{\lambda}_2$, $\widetilde{\gamma}_2$, $\widetilde{\lambda}_3$, $\widetilde{\gamma}_3$, $\widetilde{\lambda}_4$, $\widetilde{\gamma}_4$ and $\widetilde{\theta}$ of the frailty model are given in Table 3.2. The estimate for τ is 0.666. The two

models are again compared through the marginal hazards (see (3.6) for the copula model and (3.10) for the frailty model). Figure 3.2 illustrates that the marginal hazards in the copula model and in the frailty model are different. The semiparametric copula model for the infection data is given by

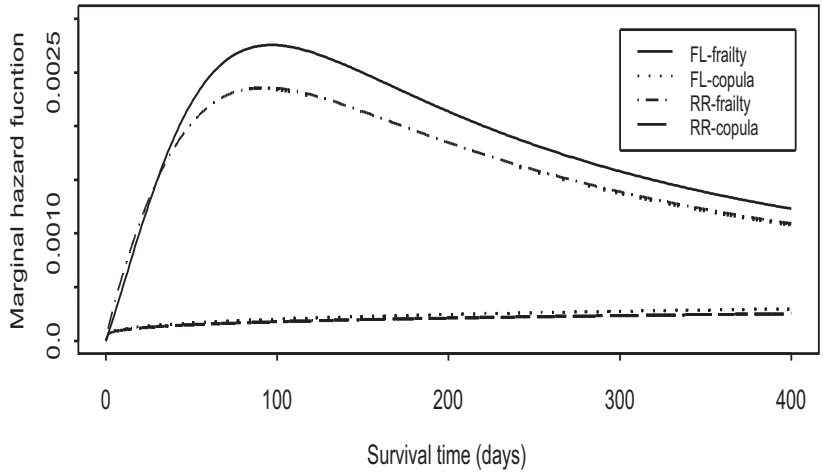

Figure 3.2: The marginal hazard functions from the frailty model and the copula model with gamma frailty density for the time to infection with *Corynebacterium bovis* for front left (FL) and rear right (RR) udder quarters.

$$h_{j,c}(t) = \begin{cases} h_0(t) & \text{for FL} \\ h_0(t)\exp(\beta_1) & \text{for RL} \\ h_0(t)\exp(\beta_2) & \text{for FR} \\ h_0(t)\exp(\beta_1+\beta_2+\beta_3) & \text{for RR} \end{cases}$$

with β_1 the effect of front versus rear, β_2 the effect of left versus right and β_3 the effect of the interaction. The nonparametric copula model is given

by

$$h_{j,c}(t) = \begin{cases} h_1(t) & \text{for FL} \\ h_2(t) & \text{for RL} \\ h_3(t) & \text{for FR} \\ h_4(t) & \text{for RR} \end{cases}$$

with $h_0(t)$, $h_1(t)$, $h_2(t)$, $h_3(t)$ and $h_4(t)$ unspecified hazard functions. Parameter estimates for the semiparametric copula model are $\hat{\beta}_1$ = -0.137 (0.038), $\hat{\beta}_2 = 0.017$ (0.035), $\hat{\beta}_3 = 0.025$ (0.052) and $\hat{\theta} = 3.386$ (0.136). The estimate for Kendall's τ is 0.629. In the nonparametric copula model $\hat{\theta}$ is equal to 3.367 (0.135), $\hat{\tau} = 0.627$.
The semiparametric frailty model is given by

$$h_{ij}(t) = \begin{cases} z_i \widetilde{h}_0(t) & \text{for FL} \\ z_i \widetilde{h}_0(t) \exp(\widetilde{\beta}_1) & \text{for RL} \\ z_i \widetilde{h}_0(t) \exp(\widetilde{\beta}_2) & \text{for FR} \\ z_i \widetilde{h}_0(t) \exp(\widetilde{\beta}_1 + \widetilde{\beta}_2 + \widetilde{\beta}_3) & \text{for RR} \end{cases}$$

with $\widetilde{\beta}_1$ the effect of front versus rear, $\widetilde{\beta}_2$ the effect of left versus right and $\widetilde{\beta}_3$ the effect of the interaction. The nonparametric frailty model is given by

$$h_{ij}(t) = \begin{cases} z_i \widetilde{h}_1(t) & \text{for FL} \\ z_i \widetilde{h}_2(t) & \text{for RL} \\ z_i \widetilde{h}_3(t) & \text{for FR} \\ z_i \widetilde{h}_4(t) & \text{for RR} \end{cases}$$

with again $\widetilde{h}_0(t)$, $\widetilde{h}_1(t)$, $\widetilde{h}_2(t)$, $\widetilde{h}_3(t)$ and $\widetilde{h}_4(t)$ unspecified hazard functions. Estimates for the parameters $\widetilde{\beta}_1$, $\widetilde{\beta}_2$, $\widetilde{\beta}_3$ and $\widetilde{\theta}$ in the semiparametric frailty model are -0.247 (0.070), 0.075 (0.068), -0.034 (0.098) and 3.872 (0.228), respectively. The estimate of Kendall's τ is 0.662. In the nonparametric frailty model the estimate for $\widetilde{\theta}$ is equal to 3.920 (0.229). The estimate of Kendall's τ is 0.662.
From the copula model as well as the frailty model it can be concluded that there is a strong positive correlation between the infection times of the four udder quarters of a cow. Both models show that the hazard of being infected is higher for front udder quarters than for rear udder quarters. No significant difference can be found between left udder quarters and right udder quarters. The estimate of β in the copula model needs to be interpreted at the marginal level: the hazard of infection of a front udder quarter of a cow is higher than the hazard of infection of a rear udder quarter of another cow. The interpretation of the estimate of $\widetilde{\beta}$ in the frailty model is at the conditional

Table 3.2: *Corynebacterium bovis* infection data: Estimates and their standard errors (Est (SE)) of the copula function parameter and the Weibull parameters of the Clayton copula model (using the two-stage approach or joint estimation) and of the shared gamma frailty model.

Parameter	Frailty model	Parameter	Copula model	
			two-stage	joint estimation
$\widetilde{\lambda}_{FL}$	0.237 (0.021)	λ_{FL}	0.148 (0.010)	0.141 (0.009)
$\widetilde{\gamma}_{FL}$	2.062 (0.069)	γ_{FL}	1.305 (0.052)	1.281 (0.048)
$\widetilde{\lambda}_{RL}$	0.195 (0.018)	λ_{RL}	0.128 (0.009)	0.121 (0.008)
$\widetilde{\gamma}_{RL}$	1.965 (0.070)	γ_{RL}	1.310 (0.055)	1.298 (0.052)
$\widetilde{\lambda}_{FR}$	0.268 (0.023)	λ_{FR}	0.157 (0.010)	0.150 (0.010)
$\widetilde{\gamma}_{FR}$	1.977 (0.065)	γ_{FR}	1.255 (0.049)	1.246 (0.046)
$\widetilde{\lambda}_{RR}$	0.213 (0.019)	λ_{RR}	0.139 (0.010)	0.128 (0.009)
$\widetilde{\gamma}_{RR}$	1.871 (0.066)	γ_{RR}	1.264 (0.052)	1.251 (0.049)
$\widetilde{\theta}$	3.991 (0.231)	θ	3.055 (0.124)	3.277 (0.186)

level: the hazard of infection of a front udder quarter of a cow is higher than the hazard of infection of a rear udder quarter of another cow with the same frailty.

3.4 The positive stable copula and frailty model

3.4.1 The diagnosis data

In the two-stage copula approach, the marginal survival functions corresponding to (3.6) are used, but the copula function now uses the Laplace transform

$$\mathcal{L}(s) = \exp\left(-s^{\theta}\right) \text{ and } \mathcal{L}^{-1}(s) = (-\log s)^{1/\theta}$$

with $0 \leqslant \theta < 1$. $\mathcal{L}(.)$ is the Laplace transform of the positive stable distribution (see (1.11)).
The corresponding copula takes the form

$$C_{\theta}(u,v) = \exp\left[-\left\{(-\log u)^{1/\theta} + (-\log v)^{1/\theta}\right\}^{\theta}\right].$$

Therefore, the joint survival function in the positive stable copula model is

$$S_c(t_1, t_2) = \exp\left[-\left[\{-\log S_{1,c}(t_1)\}^{1/\theta} + \{-\log S_{2,c}(t_2)\}^{1/\theta}\right]^{\theta}\right]. \quad (3.12)$$

The parameter estimates $\hat{\lambda}_1$, $\hat{\gamma}_1$, $\hat{\lambda}_2$ and $\hat{\gamma}_2$ are obviously the same as for the Clayton-Oakes copula model. As we did for the Clayton-Oakes copula, we replace the $S_{j,c}(.)$'s in the (joint survival functions appearing in the) likelihood and we maximize with respect to θ. Under this new dependency structure the value of θ is estimated as 0.563 (0.045). Kendall's τ in the positive stable copula is given by $\tau = 1 - \theta = 0.437$.
Since the marginal survival functions and the copula are modeled in a parametric way, the likelihood can be maximized jointly for the marginal survival function parameters and the copula function parameter. The estimates obtained from this approach are shown in Table 3.3. In the frailty model

Table 3.3: Diagnosis data: Estimates and their standard errors (Est (SE)) of the copula function parameter and the Weibull parameters of the positive stable copula model (using the two-stage approach or joint estimation) and of the positive stable frailty model.

Frailty model		Copula model		
Parameter		Parameter	two-stage	joint estimation
$\widetilde{\lambda}_{RX}$	0.020 (0.008)	λ_{RX}	0.106 (0.024)	0.118 (0.025)
$\widetilde{\gamma}_{RX}$	4.560 (0.422)	γ_{RX}	2.539 (0.191)	2.491 (0.172)
$\widetilde{\lambda}_{US}$	0.059 (0.021)	λ_{US}	0.219 (0.039)	0.213 (0.037)
$\widetilde{\gamma}_{US}$	4.240 (0.401)	γ_{US}	2.323 (0.175)	2.315 (0.180)
θ	0.546 (0.052)	θ	0.563 (0.045)	0.546 (0.052)

approach, we fit the conditional model (3.9) to the data, with the positive stable density (1.11) as frailty density. This complex expression for the positive stable density translates into the simple Laplace transform (1.12).
From (3.3) the joint survival function is

$$S_f(t_1, t_2) = \exp\left[-\{H_{1,z}(t_1) + H_{2,z}(t_2)\}^{\theta}\right].$$

Making use of (3.4) and (1.12) this can be rewritten as

$$S_f(t_1, t_2) = \exp\left[-\left[\{-\log S_{1,f}(t_1)\}^{1/\theta} + \{-\log S_{2,f}(t_2)\}^{1/\theta}\right]^{\theta}\right]$$

which has the same form as (3.12).
Also for the shared frailty model with positive stable frailty density, the frailties can be integrated out to obtain the observable likelihood which can then be maximized with respect to all the parameters (Costigan and Klein, 1993). Parameter estimates for $\widetilde{\lambda}_1$, $\widetilde{\gamma}_1$, $\widetilde{\lambda}_2$, $\widetilde{\gamma}_2$ and θ are provided in Table 3.3. For the positive stable copula with marginal Weibull hazards we have $S_{j,c}(t) = \exp\left(-\lambda_j t^{\gamma_j}\right)$; for the shared positive stable frailty model with conditional Weibull hazards $S_{j,f}(t) = \exp\left(-\widetilde{\lambda}_j^{\theta} t^{\widetilde{\gamma}_j \theta}\right)$.
So in both models the marginal survival functions are Weibull, i.e., the event times are Weibull distributed. We can make the Weibull distributions identical by taking

$$\lambda_j = \widetilde{\lambda}_j^{\theta} \qquad \gamma_j = \theta \widetilde{\gamma}_j. \tag{3.13}$$

Assuming bivariate survival data without censoring, the likelihood (which is the product over the clusters of the bivariate densities) for the copula function is

$$L_c = \prod_{i=1}^{k} \exp\left(-q_i^{\theta}\right) \lambda_1 \gamma_1 t_{i1}^{\gamma_1 - 1} \lambda_2 \gamma_2 t_{i2}^{\gamma_2 - 1} \left\{ \frac{q_i^{2(\theta-1)} + (1/\theta - 1) q_i^{\theta-2}}{(q_{i1} q_{i2})^{(\theta-1)}} \right\},$$

with $q_{ij} = \left(\lambda_j t_{ij}^{\gamma_j}\right)^{1/\theta}$ and $q_i = q_{i1} + q_{i2}$.
For the frailty model the likelihood is (after integrating out the frailties)

$$L_f = \prod_{i=1}^{k} \exp\left(-\widetilde{q}_i^{\theta}\right) \widetilde{\lambda}_1 \widetilde{\gamma}_1 t_{i1}^{\widetilde{\gamma}_1 - 1} \widetilde{\lambda}_2 \widetilde{\gamma}_2 t_{i2}^{\widetilde{\gamma}_2 - 1} \left\{ \theta^2 \widetilde{q}_i^{2(\theta-1)} + \theta(1-\theta) \widetilde{q}_i^{\theta-2} \right\},$$

with $\widetilde{q}_{ij} = \left(\widetilde{\lambda}_j t_{ij}^{\widetilde{\gamma}_j}\right)$ and $\widetilde{q}_i = \widetilde{q}_{i1} + \widetilde{q}_{i2}$.
From (3.13) we easily see that

$$\widetilde{q}_{ij} = q_{ij}. \tag{3.14}$$

Using (3.13) and (3.14) one can show that L_f can be rewritten as L_c. As an illustration check in Table 3.3 that for the estimates obtained from the maximization (jointly for all the parameters) of L_c, resp. maximization of L_f, the relations (3.13) hold.

3.4.2 The *Corynebacterium bovis* infection data

The fourdimensional joint survival function in the positive stable copula model is

$$S_c(t_1,t_2,t_3,t_4) = \exp\Big[-\Big[\{-\log S_{1,c}(t_1)\}^{1/\theta} + \{-\log S_{2,c}(t_2)\}^{1/\theta} + \{-\log S_{3,c}(t_3)\}^{1/\theta} + \{-\log S_{4,c}(t_4)\}^{1/\theta}\Big]^{\theta}\Big].$$

The parameter estimates $\hat{\lambda}_1$, $\hat{\gamma}_1$, $\hat{\lambda}_2$, $\hat{\gamma}_2$, $\hat{\lambda}_3$, $\hat{\gamma}_3$, $\hat{\lambda}_4$ and $\hat{\gamma}_4$ are the same as for the Clayton-Oakes copula model. The value of the parameter θ is estimated as 0.552 (0.009). Kendall's τ is thus 0.448.
Parameter estimates obtained by joint maximization of the likelihood for the marginal survival function parameters and the copula function parameter are shown in Table 3.4.
The joint survival function in the frailty model is given by

$$S_f(t_1,t_2,t_3,t_4) = \exp\Big[-\Big[\{-\log S_{1,f}(t_1)\}^{1/\theta} + \{-\log S_{2,f}(t_2)\}^{1/\theta} + \{-\log S_{3,f}(t_3)\}^{1/\theta} + \{-\log S_{4,f}(t_4)\}^{1/\theta}\Big]^{\theta}\Big].$$

Parameter estimates for $\widetilde{\lambda}_1$, $\widetilde{\gamma}_1$, $\widetilde{\lambda}_2$, $\widetilde{\gamma}_2$, $\widetilde{\lambda}_3$, $\widetilde{\gamma}_3$, $\widetilde{\lambda}_4$, $\widetilde{\gamma}_4$, and θ are provided in Table 3.4.
From Table 3.4 it is clear that also for the infection data the relations (3.13) hold. The fact that the parameters of the copula and frailty model can be identified, as discussed in (3.13), seems to be an exclusive property of the combination of Weibull distributed event times and frailties from a positive stable distribution. We were not able to find such relationships between the parameters for other event time distribution - frailty distribution combinations. If the exponential distribution is assumed, which is a special case of the Weibull distribution with $\gamma = 1$, for the event times together with a positive stable distribution for the frailties for example, this property no longer holds. Under these assumptions the population hazard function in the copula model $h_{j,c}(t) = \lambda_j$ is constant, but in the frailty model the marginal hazard function is no longer constant, but Weibull: $h_{j,f}(t) = \theta\widetilde{\lambda}_j^{\theta}t^{\theta-1}$.

3.5 Conclusions

In this chapter we discussed similarities and differences between copula models and frailty models for bivariate and fourdimensional survival data. We

Table 3.4: *Corynebacterium bovis* infection data: Estimates and their standard errors (Est (SE)) of the copula function parameter and the Weibull parameters of the positive stable copula model (using the two-stage approach or joint estimation) and of the positive stable frailty model.

Para - meter	Frailty model	Para- meter	Copula model two-stage	Copula model joint estimation
$\widetilde{\lambda}_{FL}$	0.025 (0.0003)	λ_{FL}	0.148 (0.010)	0.168 (0.010)
$\widetilde{\gamma}_{FL}$	2.229 (0.098)	γ_{FL}	1.305 (0.052)	1.078 (0.044)
$\widetilde{\lambda}_{RL}$	0.020 (0.0002)	λ_{RL}	0.128 (0.009)	0.152 (0.010)
$\widetilde{\gamma}_{RL}$	2.146 (0.095)	γ_{RL}	1.310 (0.055)	1.038 (0.044)
$\widetilde{\lambda}_{FR}$	0.029 (0.001)	λ_{FR}	0.157 (0.010)	0.179 (0.011)
$\widetilde{\gamma}_{FR}$	2.144 (0.021)	γ_{FR}	1.255 (0.049)	1.037 (0.042)
$\widetilde{\lambda}_{RR}$	0.022 (0.001)	λ_{RR}	0.139 (0.010)	0.159 (0.010)
$\widetilde{\gamma}_{RR}$	2.042 (0.030)	γ_{RR}	1.264 (0.052)	0.988 (0.042)
θ	0.484 (0.005)	θ	0.552 (0.009)	0.484 (0.015)

focused on the comparison between the Clayton-Oakes copula model and the shared gamma frailty model; and between the positive stable copula model and the shared positive stable frailty model. For each of the two comparisons, the copula functions used for the bivariate or fourdimensional joint survival functions are the same but the marginal survival functions are modeled in a different way. To show the differences in a concrete example, we use the Clayton-Oakes copula model with Weibull marginal survival functions and the shared gamma frailty model with conditional Weibull survival functions. A similar comparison between the positive stable copula model and the shared positive stable frailty model shows that, in the exceptional case of the Weibull hazard, there is a one-to-one match between the two models.

With the more flexible semiparametric and nonparametric model specification, parameter estimates of the copula model are typically obtained by separate modeling of the marginal survival functions (in the first stage) and the copula function (in the second stage). Therefore, there is complete separation between the estimation of the marginal survival function parameters and the copula function parameter. In the frailty model however, the frailty parameter appears in the marginal survival functions, making separate estimation impossible.

From a practical point of view, the choice between the two models depends on different considerations. First of all, the data structure is a limiting factor. Copula models require small and equal cluster sizes. Data sets with large and/or unequal cluster sizes are however frequently encountered and can be handled by the frailty model. The interpretation of covariate effects might be more straightforward in the copula model for a non-statistician. Covariate effects need to be interpreted at the conditional level in the frailty model, while interpretation of the covariate effects is at the marginal level in the copula model.
For the diagnosis data it can be concluded that the time to diagnosis of fracture healing in dogs can be shortened by using the ultrasound technique and that the use of models that take into account the correlation in the data was necessary because of existing positive correlation in the data. Considering that the US technique is cheaper and that there would be no roentgen exposure of dogs and staff, diagnosis with US is a promising technique. Earlier diagnosis of a healed fracture by US can prevent unnecessarily long limb immobilization and allow earlier dynamization.

Chapter 4

An overview of current methods for interval-censored data

Based on:
Goethals, K., Janssen, P., and Duchateau, L. (2011), "Parametric estimation in the copula model for fourdimensional interval-censored failure time data," *in preparation*

4.1 Introduction

In this chapter we start with a short overview of available nonparametric, parametric and semiparametric methods to analyze univariate interval-censored data in Section 4.2. Next, some statistical techniques, available in commercial software packages, to model multivariate interval-censored data are discussed in Section 4.3. The advantages and disadvantages of the discussed marginal, fixed effects and copula model are illustrated using the mastitis data described in Section 1.9.2. A major drawback of the discussed copula model is the necessity to use a two-step procedure in which midpoint imputation in the second step is required because only the likelihood for right-censored data is available in the literature. This problem is solved in Section 4.4 in which we describe the construction of the likelihood for fourdimensional interval-censored data in the copula model.

4.2 Univariate interval-censored data

One approach towards modeling interval-censored data is to reduce the problem of analyzing interval-censored event time data to analyzing right-censored event time data, called the imputation approach. That way existing nonparametric, semiparametric and parametric inference procedures and statistical software developed for right-censored data can be used. Two imputation approaches are generally used: single point imputation and multiple imputation. Single point imputation is commonly used in practice for its simplicity. In single point imputation it is assumed that the underlying true event time is equal to a value within the observed interval. Common choices for that value are the midpoint, the upper bound or the lower bound of the interval. If the intervals are narrow, the three choices will not give very different results. Another option is to randomly select a value in the observed interval. The multiple imputation approach is based on the data augmentation algorithm given in Tanner and Wong (1987) and iterates between an imputation and an estimation step until convergence. Authors that considered the multiple imputation approach include Satten et al. (1998), Bebchuk and Betensky (2000) and Pan (2000). However, the imputation approach may lead to biased estimates if the intervals are wide and varying (Odell et al., 1992; Goggins et al., 1998). Furthermore, the standard errors of the coefficients will be underestimated since the event times are treated as known when, actually, they are not (Goggins et al., 1998). In the next subsections we discuss some specific methods that explicitly model the interval

censoring in the data.

4.2.1 Nonparametric methods

For right-censored data, the nonparametric maximum likelihood estimate (NPMLE) of the survival function is given by Kaplan and Meier (1958). In the case of interval-censored data the survival function can also be estimated nonparametrically, however, nonparametric inference is much more complicated for interval-censored data. In general, the NPMLE of the survival function does not have a closed form and can only be determined using iterative algorithms. The first NPMLE for interval-censored data was proposed by Peto (1973) using the constrained Newton-Raphson method. However, the most commonly known NPMLE for interval-censored data is Turnbull's estimate (Turnbull, 1976), who developed the self-consistency algorithm, which is an EM-like algorithm (Dempster et al., 1977). Other algorithms to maximize the likelihood function are the iterative convex minorant (ICM) algorithm, first introduced by Groeneboom and Wellner (1992) and later modified by Jongbloed (1998) and the EM-ICM algorithm, proposed by Wellner and Zhan (1997), which combines the self-consistency algorithm and the ICM algorithm.
Sometimes a smooth estimate of the survival function is more desirable than Turnbull's step-wise estimate. One way of smoothing the survival or equivalently the density function is proposed by Kooperberg and Stone (1992), who smooth the density function using splines. Figure 4.1 shows the nonparametric estimate of Turnbull and the smooth estimate of Kooperberg and Stone of the survival function for the interval-censored time to infection for the left front udder quarters of primiparous cows. Other approaches to obtain smooth estimates include kernel-based methods and penalized or local likelihood methods (Sun, 2006).

4.2.2 Parametric methods

If only interval-censored data of the form $\lfloor L_i, U_i \rceil$; $i = 1, \ldots, n$ are available, and if noninformative censoring is assumed, the likelihood function is proportional to

$$L(\boldsymbol{\zeta}) \approx \prod_{i=1}^{n} \left[S(l_i) - S(u_i) \right], \tag{4.1}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard. If we assume a parametric form for the survival function $S(.)$ standard maximum likelihood theory can be applied to obtain parameter estimates. The main

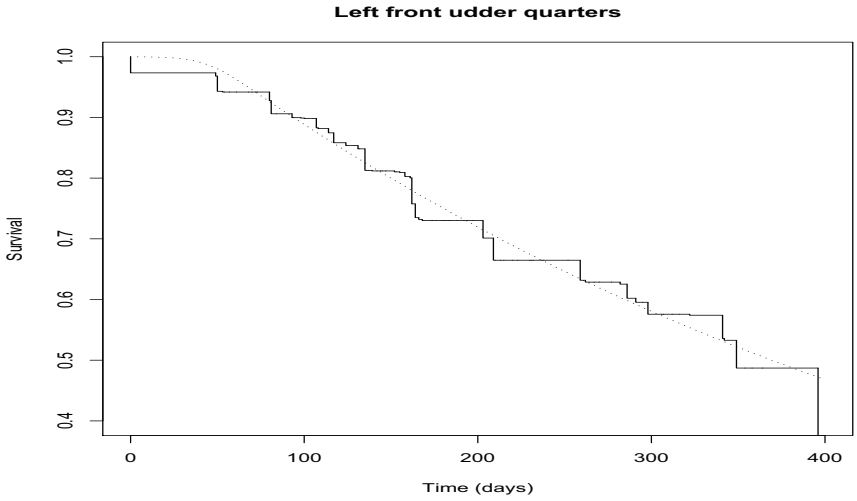

Figure 4.1: Survival function estimates of the interval-censored time to infection for the left front udder quarters of primiparous cows. Nonparametric estimate of Turnbull (solid line), smooth estimate of Kooperberg and Stone (dotted line).

advantage of parametric methods is that their implementation is straightforward and available in commercial software packages.

4.2.3 Semiparametric methods

The ordinary partial likelihood maximization procedure for right-censored data can not be used in the case of interval-censored data. Several authors propose methods to analyze interval-censored data semiparametrically, but the proposed methods are computationally demanding and not available in commercial software packages. Finkelstein (1986) and Goetghebeur and Ryan (2000) propose a full likelihood approach in which the baseline hazard is estimated nonparametrically simultaneously with the regression coefficients. The method of Finkelstein (1986) is based on the grouped data assumption, but the method of Goetghebeur and Ryan (2000) relaxes this assumption. Satten (1996) and Goggins et al. (1998) investigate a marginal likelihood approach based on the likelihood given by the sum over all rankings of the underlying and unobserved failure times that are consistent with the observed censoring intervals. This approach focuses only on the re-

gression coefficients. Betensky et al. (2002) and Cai and Betensky (2003) consider methods in which the baseline hazard is approximated by finite-dimensional functions.

4.3 Multivariate interval-censored data

For the analysis of multivariate interval-censored data, one has to deal simultaneously with the interval censoring problem and the clustering. To simplify the analysis either the interval-censored nature of the data or the correlation structure in the data could be ignored. For instance, when ignoring the interval censoring, the imputation approach, in which the midpoint, lower or upper bound of the interval is imputed as an exact event time, can be used and standard analytic techniques for multivariate survival data can be applied (Wei and Glidden, 1997; Hougaard, 2000; Kelly and Lim, 2000). However, as mentioned before, this approach may lead to biased estimates if the intervals are wide and varying (Odell et al., 1992; Goggins et al., 1998) and the standard errors of the coefficients will be underestimated (Goggins et al., 1998). Another alternative is to ignore the correlation between the observations and to analyze the data using methods for univariate interval-censored data (see Section 4.2). However, it is well known that clustering in the data needs to be accounted for in the analysis, otherwise inferences and standard errors will not be correct (Wei and Glidden, 1997).
Therefore, it would be best to use models that can deal with the clustering and interval censoring simultaneously. In this section we give an overview of statistical techniques to model multivariate interval-censored survival data available in commercial software packages. In the first subsection the marginal model, the fixed effects model and the two-stage copula model will be discussed. In the next subsection an overview is given of three commercial software packages that can fit these models and the last subsection describes the results obtained when applying these models to the mastitis data. Focus is on parametric frequentist approaches, Bayesian and semi- or nonparametric frequentist methods are not discussed.

4.3.1 The models

If there is no interest in the correlation parameter, a marginal model or a fixed effects model for interval-censored data could be used for the analysis. However, techniques to model data which are simultaneously clustered and interval-censored and provide an estimate of the correlation received less attention in the literature. Possible options are the frailty model and the

copula model. For a discussion on frailty models for interval-censored data we refer to Chapter 5. The copula model is discussed in this subsection and in the next section.

The marginal model

In the marginal model approach the clustering is not taken into account and the event times are treated as if they were independent of each other even if they belong to the same cluster. In this model the only difference between subjects is their covariate information. Therefore, parameter estimates will refer to a randomly selected subject from the population, not taking into account which cluster the subject belongs to. That is why these models are also called population-averaged models. In the proportional hazards formulation the model is given by

$$h_{ij}(t) = h_0(t) \exp(\mathbf{x}_{ij}^t \boldsymbol{\beta}), \tag{4.2}$$

with $h_{ij}(t)$ the hazard at time t for subject j of cluster i ($i = 1, \ldots, k$), $h_0(t)$ the baseline hazard, $\mathbf{x}_{ij}$ the vector of covariates for the corresponding subject and $\boldsymbol{\beta}$ the vector of covariate effects.
Since the existing correlation between subjects is ignored, one might be concerned with the consistency of the parameter estimates. Bogaerts et al. (2002) proof that, even though the correlation between the observations is ignored, these estimates are consistent under a correct parametric specification, but the likelihood-based estimates of their variance, the inverse of the information matrix, are not because of the ignored correlation between event times. Following the approach of Royall (1986), Bogaerts et al. (2002) derive consistent estimators for the variance under model misspecification for interval-censored multivariate survival data. Another way to obtain a consistent estimate of the variance is to use the grouped jackknife technique (Lipsitz et al., 1994). Denote the p-dimensional vector of parameter estimates using the whole data set by $\boldsymbol{\zeta}$. We now leave out each of the k clusters of observations one by one and fit model (4.2) to each new data set, resulting for the data set with cluster i deleted, in parameter vector estimate $\hat{\boldsymbol{\zeta}}_{-i}$, $i = 1 \ldots, k$. The grouped jackknife estimate is then given by (Wu, 1986)

$$\left(\frac{k-p}{k}\right) \sum_{i=1}^{k} \left(\hat{\boldsymbol{\zeta}}_{-i} - \hat{\boldsymbol{\zeta}}\right) \left(\hat{\boldsymbol{\zeta}}_{-i} - \hat{\boldsymbol{\zeta}}\right)^t. \tag{4.3}$$

We will use the grouped jackknife technique to obtain estimates for the variance when analyzing the mastitis data with the marginal model.

The fixed effects model

One way of actually modeling the cluster effect is to add a fixed cluster effect to model (4.2). The hazard is now given by

$$h_{ij}(t) = h_0(t) \exp(\mathbf{x}_{ij}^t \boldsymbol{\beta} + c_i),$$

with c_i the fixed effect for the i^{th} cluster. This model is overparameterized. Therefore, the restriction $c_i = 0$ for a certain i $(i = 1, \ldots, k)$ is added. An estimate and variance for each fixed cluster effect is provided which is not really of interest; we merely want to adjust the model to account for the clustering in the data.

The copula model

None of the models discussed above gives an actual estimate of the clustering effect in the data. If the correlation between the observations itself is of interest the copula model can be used to obtain an estimate of the strength of the clustering. For a description of the copula model, see Section 1.8. In this chapter we assume a Clayton-Oakes copula model (see 3.11). In the first stage of the estimation procedure (see Section 1.8), the marginal model (4.2) is fit to the data, using the interval-censored nature of the data. In the second stage, on the other hand, the midpoint imputation approach is used since the likelihood for interval-censored data is not available in the literature. The likelihood for right-censored data that then needs to be optimized to obtain an estimate for θ is given in Massonnet et al. (2009).

4.3.2 Software

While we have discussed the different model types in terms of the proportional hazards formulation (see Section 1.5.1), the software packages SAS, S-Plus and R discussed in this section use the loglinear model representation (see Section 1.5.3). To transform the parameters obtained in the software's output to the proportional hazards formulation, we will make use of formulas (1.6) to (1.8). The software package STATA fits the models in the proportional hazards formulation.

SAS

The current version of SAS is SAS 9.2. Parametric survival models for event time data that can be right-, left- or interval-censored can be fitted by using

the `proc lifereg` procedure. All time points should be recorded as an interval. If the lower bound of the interval is missing, the upper bound is used as a left-censored value. Similarly, if the upper bound is missing, the lower bound is taken as a right-censored value. A censoring interval with the same endpoints is considered to be an exact event time, otherwise it represents an interval-censored observation. Possible survival distributions are exponential, Weibull, lognormal, loglogistic, normal, logistic and generalized gamma.

The marginal model. The option `covsandwich (aggregate)` to obtain a robust variance estimate is only available in the `proc phreg` procedure and not in the `proc lifereg` procedure. The user is forced to implement the grouped jackknife technique by writing a loop in which the clusters of observations are left out one by one, the model is fitted k times and then using formula (4.3).

The fixed effects model. The user has to declare the cluster variable as categorical using the `class` statement. When fitting an overparameterized model the coefficient of the effect with the largest alphanumeric value is put equal to zero. The `proc lifereg` procedure gives us an estimate of σ and its variance which enables us to calculate the variances of $\hat{\lambda}$ and $\hat{\beta}$.

The copula model. Estimates for the baseline hazard parameters and the covariate effects in the first stage of the estimation procedure for the copula model are obtained by fitting a marginal model through the `proc lifereg` procedure. To obtain an estimate for θ the user is forced to write his own program.

S-Plus and R

Both S-Plus and R use the S language and share many of the same functions. R is a free software package and can be downloaded from http://www.r-project.org. Current versions are S-Plus 8.2 and R 2.13.0.
The `censorReg` or `survReg` function in S-Plus or the `survreg` function in R fits a parametric survival model to arbitrarily censored event time data. In the `censorReg` function the response is usually an object of class `censor` as computed by the `censor` function. The response for the `survReg/survreg` function in S-Plus/R is usually an object of class `surv` as computed by the `surv` function. The type of censoring is indicated by the status indicator and should be 0 for right-censored data, 1 for an exact event time, 2 for left-censored data and 3 for interval-censored data. In case of a right-censored

observation the `censor` or `surv` function takes the lower bound of the interval as time point. The time point for a left-censored observation is the upper bound of the interval. Possible survival distributions are exponential, Weibull, normal, lognormal, logistic, loglogistic, extreme value, Rayleigh and t. It is important to note that in the `censorReg` function the exponential distribution is not defined as the Weibull distribution with the scale parameter fixed to one as in the `survR(r)eg` function, but as the minimum extreme value distribution with the scale parameter fixed to one.

The marginal model. A robust variance estimator can be obtained in Cox models using the `cluster(id)` option in `coxph`. If the values in `id` are not unique, but instead identify clusters of correlated observations, then the variance estimate is based on the grouped jackknife technique. This is not an option in `SurvR(r)eg` or `censorReg`. The grouped jackknife technique as described in Section 4.3.1 can however be implemented by the user in S-Plus or R by writing a loop leaving out the clusters of observations one by one, fitting the model k times and using formula (4.3).

The fixed effects model. The function `as.factor` is used to interpret the cluster variable as a categorical variable. When fitting the overparameterized fixed effects model, S-Plus and R put the coefficient of the fixed effect that starts with the lowest alphanumeric value to zero. Since the `censorReg` and `survR(r)eg` functions provide us only with the estimate and its variance of $\log(\sigma)$, $\hat{\sigma}$ needs to be calculated as $\exp(\log(\hat{\sigma}))$ and its variance is $\hat{\sigma}^2\text{var}(\log(\hat{\sigma}))$. An estimate and corresponding variance of γ can be obtained using (1.5) and (1.7), but it is not possible to calculate the variances of $\hat{\lambda}$ or $\hat{\beta}$ since only the covariance of $\hat{\mu}$ and $\log(\hat{\sigma})$ or $\hat{\alpha}$ and $\log(\hat{\sigma})$ is given instead of the covariance of these parameters and $\hat{\sigma}$. Obviously this is a major drawback of the `censorReg` and `survR(r)eg` functions.

The copula model. Estimates for the baseline hazard parameters and the covariate effects in the first stage of the estimation procedure for the copula model are obtained by fitting a marginal model through the `censorReg` or `survR(r)eg` functions. The second stage of the estimation procedure needs to be implemented by the user himself.

STATA

Contrary to SAS and S-Plus/R that provide parameter estimates of the log-linear model representation, Stata fits the model in the proportional hazards

formulation unless the option `time` is specified. The option `time` is only valid for the exponential and Weibull models since they have both a proportional hazards and an accelerated failure time parameterization. `Intcens` is a Stata module written by Jamie Griffin (Griffin, 2005) that performs a parametric interval-censored survival analysis. The program fits various distributions by maximum likelihood to non-negative data which can be left-, right- or interval-censored. The supported distributions are exponential, Weibull, Gompertz, log-logistic, log-normal, 2 and 3 parameter gamma and inverse Gaussian. This module requires a Stata version 8.2 or later, current version is Stata 11. Before performing the analysis you have to declare your data to be event time data by the `stset` command with the `failure(failvar)` option. If the failure option is not specified, all records are assumed to end in failure. If it is specified, `failvar` is interpreted as an indicator variable; 0 and missing mean censored, and all other values are interpreted as representing failure. It is also possible to define your own list of values that indicate failure (StataCorp., 2005).

The marginal model. The option `robust` indicates that the sandwich estimator of variance needs to be used instead of the traditional calculation. Adding the option `cluster(varname)`, where `varname` specifies to which group each observation belongs, states that the observations are independent across groups (clusters), but not necessarily within groups.

The fixed effects model. The user needs to specify the `xi` option to indicate that the cluster variable is categorical. The coefficient of the effect that starts with the smallest alphanumeric value is put equal to zero, this can be seen in the output. `Intcens` gives us an estimate of γ and β and its standard error. An estimate for λ needs to be calculated as $\exp(\hat{\nu})$ where $\hat{\nu}$ is given in the output as `const` and its variance is $(\exp(\hat{\nu}))^2 \operatorname{var}(\hat{\nu})$.

The copula model. Estimates for the baseline hazard parameters and the covariate effects in the first stage of the estimation procedure for the copula model are obtained by fitting a marginal model through the `Intcens` module. The user is forced to implement the second stage himself.

4.3.3 Analysis of the mastitis data

The mastitis data set (see Section 1.9.2) is an example of a data set that is simultaneously clustered and interval-censored: the udder quarter infection times are clustered within the cow udder and the udder quarter infection

status is followed up only periodically. However, the information available in udder quarter infection data is often not exploited to its full extent and the specific data structure is not always modeled correctly. Because of the complexity of interval-censored data, udder quarter infection data have often been reduced to binary data (Schukken et al., 1999; Zadoks et al., 2001): either an infection occurs in the udder quarter or not during the complete lactation period. However, this reduces the amount of information considerably.

We investigate the effect of the udder quarter location (front or rear), an udder quarter level covariate, and the effect of parity (multiparous versus primiparous), a between cow covariate, on time to infection with four different bacteria, i.e. *Staphylococcus aureus* (*Staph. aureus*), *Corynebacterium bovis* (*C. bovis*), *Streptococcus dysgalactiae* (*Strep. dysgalactiae*) and *Streptococcus uberis* (*Strep. uberis*). We assume a Weibull distribution for the baseline hazard. The likelihood that needs to be maximized in the marginal model and the fixed effects model for the mastitis data is an extension of expression (4.1) to data that can be interval-censored and right-censored and is given by

$$L(\boldsymbol{\zeta}) \approx \prod_{i=1}^{n} [S(l_i) - S(u_i)]^{\delta_i} \, S(l_i)^{1-\delta_i},$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard. The parameter estimates and their standard errors for the four bacteria are given in Table 4.1 for the different models.

In the marginal model we obtain the following results. The rear udder quarters have a (marginally) significant higher hazard of infection than the front udder quarters for *Staph. aureus*, with hazard ratio (= $\exp(\hat{\beta}_l)$) (HR) = 1.33 (95% confidence interval (CI) [1.03;1.70]), but a significantly lower hazard rate for the rear udder quarters is observed for *C. bovis*, with HR = 0.88 (95% CI [0.83;0.93]). For the two other bacteria, no significant differences were found, with the hazard ratio equal to 1.40 (95% CI [0.93;2.11] (marginally insignificant)) for *Strep. dysgalactiae* and equal to 1.19 (95% CI [0.92;1.54] (marginally insignificant)) for *Strep. uberis*. The hazard of infection for multiparous cows was significantly higher compared to heifers for *C. bovis* (HR = $\exp(\hat{\beta}_p)$=1.52, 95% CI [1.28;1.80]) and *Strep. uberis* (HR = 2.18, 95% CI [1.40;3.38]). The hazard of infection for multiparous cows was also higher compared to heifers for *Staph. aureus* (HR = 1.23, 95% CI [0.87;1.75]) and *Strep. dysgalactiae* (HR = 1.12, 95% CI [0.69;1.80]), but not significantly. For the derivation of the 95% confidence intervals, it is

required to use the jackknife estimate of the standard error. Both the naive standard error estimate and the jackknife estimate are given in Table 4.1. It is apparent that the jackknife estimate of the standard error can be either lower, as is the case for the udder quarter location covariate, or higher, as is the case for the parity covariate, than the naive standard error estimate. In the marginal model the hazard ratio represents the hazard of infection for a randomly chosen rear udder quarter versus the hazard of infection for a randomly chosen front udder quarter from whatever other cow.
In the copula model the estimates of the marginal model can be used to obtain the marginal survival functions, required as input in the copula model. The copula model estimate for θ is equal to 5.151 for *Staph. aureus*, 2.971 for *C. bovis*, 3.048 for *Strep. dysgalactiae* and 5.273 for *Strep. uberis*, leading to values for Kendall's τ of 0.72, 0.60, 0.60 and 0.73, respectively. Infection times within a cow are highly correlated for all four bacteria. The hazard ratio needs to be interpreted in the same way as in the marginal model.
The fixed effects model was also fitted to the data. This model has to be used with caution for the type of clustered survival data presented here. To start with, the interpretation of the parameter λ is quite different. As the fixed effect of the first cow has been put to zero, the parameter estimate for λ actually corresponds to the parameter for the first cow only, its hazard being equal to $\lambda \exp(x_{ij}\beta)\gamma t^{\gamma-1}$. As the first cow did not have any infections, it is actually impossible to estimate this parameter, therefore, the estimate is typically put at an arbitrary low value with a high standard error. Therefore, with the hazard given for any other cow i by $\lambda \exp(x_{ij}\beta + c_i)\gamma t^{\gamma-1}$, the estimate of the fixed cow effect is large for cows with infections to counteract the effect of the small value for the estimate of λ , and also these fixed effects have large standard errors. The hazard ratio for rear udder quarters versus front udder quarters is given by 1.53 (95% CI [1.27;1.84]), 0.71 (95% CI [0.65;0.79]), 1.55 (95% CI [1.17;2.05]) and 1.22 (95% CI [0.99;1.50]) for *Staph. aureus*, *C. bovis*, *Strep. dysgalactiae* and *Strep. uberis*, respectively. Contrary to the marginal and copula model, the effect of the udder quarter location covariate is also significant for infection with *Strep. dysgalactiae*. The hazard ratio for multiparous versus heifer cannot be obtained in the fixed effects model, because there is complete confounding between the cow fixed effects and the parity covariate, in the sense that the parity covariate can be written as a linear function of the cow fixed effects. Nevertheless, statistical software packages typically provide an estimate for this hazard ratio. For instance, if the parity covariate is introduced first in the model (and therefore not adjusted for the fixed cow effects), an impossibly high estimate for β_p equal to -23 results, with the HR given by exp(-23). On

Table 4.1: Parameter estimates (Est) and their standard errors (SE) for the marginal model, the fixed effects model and the copula model with parity (with β_p the effect of a multiparous cow) and udder quarter location (with β_l the effect of a rear udder quarter) as covariates and Weibull baseline hazard for infection with either *Staphylococcus aureus*, *Corynebacterium bovis*, *Streptococcus dysgalactiae* or *Streptococcus uberis*.

Bacterium	Parameter	Marginal Est (Naive/Jackknife SE)	Fixed effects Est (SE)	Copula Est (Jackknife SE)
Staphylococcus aureus	θ	-	-	5.151
	λ	0.013 (0.002/0.002)	-	0.013 (0.002)
	γ	0.996 (0.063/0.049)	1.301 (0.052)	0.996 (0.049)
	β_l	0.285 (0.133/0.124)	0.424 (0.095)	0.285 (0.124)
	β_p	0.210 (0.144/0.179)	-	0.210 (0.179)
Corynebacterium bovis	θ	-	-	2.971
	λ	0.115 (0.006/0.009)	-	0.115 (0.009)
	γ	1.285 (0.027/0.031)	2.345 (0.049)	1.285 (0.031)
	β_l	-0.129 (0.046/0.030)	-0.337 (0.049)	-0.129 (0.030)
	β_p	0.417 (0.051/0.086)	-	0.417 (0.086)
Streptococcus dysgalactiae	θ	-	-	3.048
	λ	0.005 (0.001/0.001)	-	0.005 (0.001)
	γ	1.008 (0.104/0.057)	1.301 (0.052)	1.008 (0.057)
	β_l	0.335 (0.219/0.209)	0.424 (0.095)	0.335 (0.209)
	β_p	0.110 (0.133/0.245)	-	0.110 (0.245)
Streptococcus uberis	θ	-	-	5.273
	λ	0.007 (0.002/0.001)	-	0.007 (0.001)
	γ	1.047 (0.075/0.061)	1.248 (0.087)	1.047 (0.061)
	β_l	0.173 (0.150/0.132)	0.201 (0.105)	0.173 (0.132)
	β_p	0.779 (0.194/0.224)	-	0.779 (0.224)

the other hand, if the parity covariate is introduced in the model after the fixed cow effects (and therefore adjusted), the estimate for β_p equals 0, with the HR given by 1. Both results are obviously meaningless. The fixed effects model is a conditional model. The hazard ratio reflects the hazard of infection of a rear versus a front udder quarter within the same cow.

4.3.4 Conclusions

The use of the fixed effects model is discouraged for the mastitis data and other (infection) data consisting of many small clusters. The high number of clusters that need to be added as fixed effects often causes computational problems in the software packages due to insufficient memory making it impossible to fit the fixed effects model. If the fitting of the model is successful, parameter estimates can be strange and interpretation should be careful. For example, the parameter λ of the Weibull distribution only corresponds to the first cow. If the observations of all the udder quarters of this cow are censored, this parameter can not be estimated. However, most software packages still provide a (senseless) estimate for this parameter and its standard error. It is further impossible to obtain parameter estimates for cow level covariates because there is complete confounding between the cow fixed effects and the cow level covariates. But still, statistical software packages provide a senseless estimate for these effects. Finally, estimable fixed effects need to be interpreted at a conditional level.
A more valuable alternative for the fixed effects model is the marginal model. Parameter estimates are consistent and consistent estimators of the variance are available or the grouped jackknife technique can be used to obtain consistent variance estimates. However, the techniques to obtain consistent variance estimates are not available in SAS or SPlus/R. The use of correct variances is important in order to draw correct conclusions concerning the covariate effects. The marginal model does not provide an estimate of the strength of the clustering in the data, but if interest is only in the covariate effects, the use of the marginal model is recommended. It is applicable for data sets consisting of variable cluster sizes and the number of clusters is no restriction. Finally, the interpretation of the covariate effects is at the marginal level.
If interest is also in the strength of the clustering in the data, the copula model should be used. Copula models can only be used when cluster sizes are equal and preferably small. Another important drawback of copula models for interval-censored data at the moment is the necessity to use a two-stage procedure. In the first stage the interval-censored nature of the

data is taken into account when fitting a marginal model to the data, but in the second stage imputation of, for example, the midpoint as an exact event time is necessary because only the likelihood for right-censored data is available in the literature (Massonnet et al., 2009). In the next section this problem is solved by describing the construction of the likelihood for the fourdimensional copula model for interval-censored data.

4.4 A fourdimensional copula model for interval-censored data

In the previous section a copula model is fitted to fourdimensional interval-censored data by a two-stage procedure. In the first stage, a univariate marginal model for interval-censored data is used to obtain estimates for the marginal survival functions. In the second stage however, the estimated marginal survival functions are plugged in a likelihood expression for right-censored data. The midpoint of the interval is then used as an exact event time in case of an event. This method only accounts for the interval-censored nature of the data in the first stage.
Sun et al. (2006) propose a two-stage estimation procedure to estimate the correlation when bivariate interval-censored event time data are available. In both stages the interval-censored nature of the data is accounted for. In the first stage the marginal survival functions are estimated nonparametrically and then θ is estimated by maximizing the loglikelihood with the marginal survival functions replaced by their estimates in the second stage. They proof that under certain regularity conditions $\hat{\theta}$ is consistent and asymptotically normal. They give a consistent estimator for the variance of $\hat{\theta}$, but since the calculation of this estimator could be very technically involved they propose to use the bootstrap procedure for variance estimation.
In this section we will extend their approach to the fourdimensional case, but since we only consider parametric marginal survival functions, a one-stage estimation procedure can be used.

4.4.1 Construction of the likelihood

Bivariate data

For bivariate data that can be right- or interval-censored the likelihood consists of four different possible contributions, depending on the censoring status of the two subjects in the cluster (Sun et al., 2006). A cluster with

two interval-censored observations has contribution

$$\begin{aligned} L_{i,(1,1)} &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}\right) \\ &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}\right) - P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant u_{i2}\right) \\ &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}\right) \\ &\quad -P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}\right) + P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant u_{i2}\right) \\ &= S(l_{i1}, l_{i2}) - S(u_{i1}, l_{i2}) - S(l_{i1}, u_{i2}) + S(u_{i1}, u_{i2}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2})\right) - C_\theta\left(S_1(u_{i1}), S_2(l_{i2})\right) \\ &\quad -C_\theta\left(S_1(l_{i1}), S_2(u_{i2})\right) + C_\theta\left(S_1(u_{i1}), S_2(u_{i2})\right). \end{aligned}$$

Figure 4.2 depicts $P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}\right)$ in a twodimensional plot. Figure 4.3 shows how this chance is constructed. The top left panel represents $P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}\right)$, the top right panel illustrates the subtraction of $P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}\right)$, the bottom left panel shows the subtraction of $P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}\right)$ and the bottom right panel illustrates that $P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant u_{i2}\right)$ needs to be added again since this part was sub-

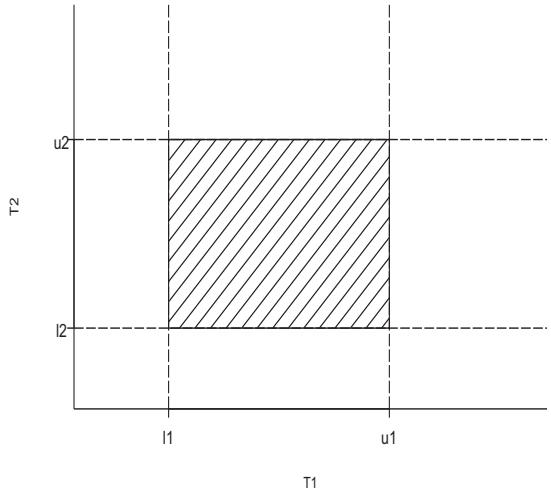

Figure 4.2: Visual representation of the likelihood contribution of a cluster with two interval-censored observations for the copula model in a twodimensional plot.

tracted twice. The contribution of a cluster where the first observation is

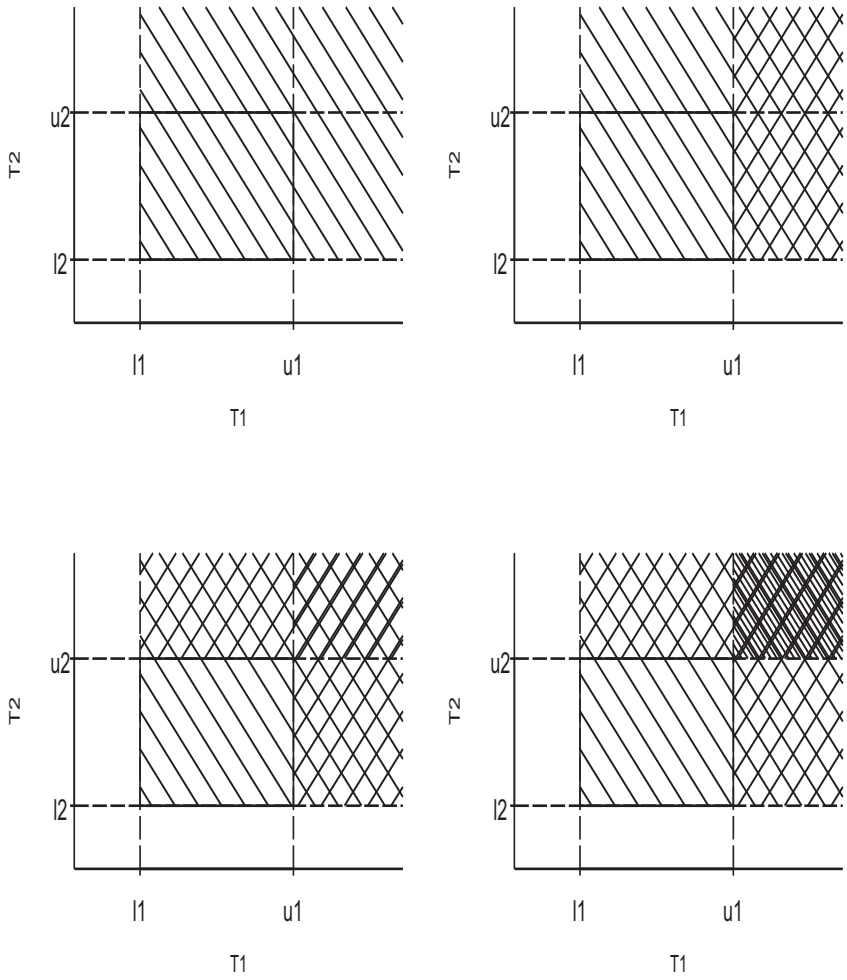

Figure 4.3: Visual representation of the construction of the likelihood contribution of a cluster with two interval-censored observations for the copula model in a twodimensional plot.

interval-censored and the second observation is right-censored is

$$\begin{aligned} L_{i,(1,0)} &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}\right) \\ &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}\right) \\ &= S(l_{i1}, l_{i2}) - S(u_{i1}, l_{i2}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2})\right) - C_\theta\left(S_1(u_{i1}), S_2(l_{i2})\right). \end{aligned}$$

The contribution of a cluster where the first observation is right-censored and the second observation is interval-censored is

$$\begin{aligned} L_{i,(0,1)} &= P\left(T_{i1} \geqslant l_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}\right) \\ &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}\right) - P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}\right) \\ &= S(l_{i1}, l_{i2}) - S(l_{i1}, u_{i2}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2})\right) - C_\theta\left(S_1(l_{i1}), S_2(u_{i2})\right). \end{aligned}$$

A cluster with two right-censored observations has contribution

$$\begin{aligned} L_{i,(0,0)} &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}\right) \\ &= S(l_{i1}, l_{i2}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2})\right). \end{aligned}$$

In the contributions above $S(r, s)$ is the joint marginal survival function, the lower bound of the interval is used as the censoring time for a right-censored observation. In the copula model the joint marginal survival function is a copula function $C_\theta(v_1, v_2)$ with arguments (v_1, v_2) the marginal survival functions $S_1(r)$ and $S_2(s)$.
The loglikelihood is then given by

$$\begin{aligned} \log L(\boldsymbol{\zeta}) = \sum_{i=1}^{k} \Big[& (1-\delta_{i1})(1-\delta_{i2}) \log L_{i(0,0)}(l_{i1}, l_{i2}) \\ & + \delta_{i1}(1-\delta_{i2}) \log L_{i(1,0)}(l_{i1}, l_{i2}, u_{i1}) \\ & + (1-\delta_{i1})\delta_{i2} \log L_{i(0,1)}(l_{i1}, l_{i2}, u_{i2}) \\ & + \delta_{i1}\delta_{i2} \log L_{i(1,1)}(l_{i1}, l_{i2}, u_{i1}, u_{i2}) \Big], \end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard, k the number of clusters, $\delta_{ij}, j = 1, 2$ equal to 1 if the j^{th} observation of the i^{th} pair is interval-censored and equal to 0 if the j^{th} observation of the i^{th} pair is right-censored.

Trivariate data

We will first describe the construction of the different likelihood contributions for the trivariate copula model for interval-censored data in order to keep formulae surveyable. The construction of the different likelihood contributions for the fourdimensional copula model for interval-censored data can be obtained in a similar way. The likelihood for trivariate data that can be right- or interval-censored consists of eight different possible contributions, depending on the censoring status of the three subjects in the cluster. A cluster with three interval-censored observations has contribution

$$\begin{aligned} L_{i,(1,1,1)} &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}, l_{i3} \leqslant T_{i3} \leqslant u_{i3}\right) \\ &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &\quad -P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}, T_{i3} \geqslant u_{i3}\right). \end{aligned} \tag{4.4}$$

The first term in the right hand side of (4.4) is equal to

$$\begin{aligned} &P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ -&P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right), \end{aligned}$$

which is equal to

$$\begin{aligned} &P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ -&P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right) + P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right). \end{aligned}$$

The second term in the right hand side of (4.4) is equal to

$$\begin{aligned} &P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant u_{i3}\right) \\ -&P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant u_{i3}\right), \end{aligned}$$

which is equal to

$$\begin{aligned} &P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant u_{i3}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant u_{i3}\right) \\ -&P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant u_{i3}\right) + P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant u_{i3}\right). \end{aligned}$$

Bringing all the terms together, the contribution for a cluster with three interval-censored observations is

$$\begin{aligned} L_{i,(1,1,1)} = &S(l_{i1}, l_{i2}, l_{i3}) - S(u_{i1}, l_{i2}, l_{i3}) - S(l_{i1}, u_{i2}, l_{i3}) + S(u_{i1}, u_{i2}, l_{i3}) \\ &-S(l_{i1}, l_{i2}, u_{i3}) + S(u_{i1}, l_{i2}, u_{i3}) + S(l_{i1}, u_{i2}, u_{i3}) - S(u_{i1}, u_{i2}, u_{i3}) \end{aligned}$$

or in terms of the copula

$$\begin{aligned} L_{i,(1,1,1)} &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2}), S_3(l_{i3})\right) - C_\theta\left(S_1(u_{i1}), S_2(l_{i2}), S_3(l_{i3})\right) \\ &\quad -C_\theta\left(S_1(l_{i1}), S_2(u_{i2}), S_3(l_{i3})\right) + C_\theta\left(S_1(u_{i1}), S_2(u_{i2}), S_3(l_{i3})\right) \\ &\quad -C_\theta\left(S_1(l_{i1}), S_2(l_{i2}), S_3(u_{i3})\right) + C_\theta\left(S_1(u_{i1}), S_2(l_{i2}), S_3(u_{i3})\right) \\ &\quad +C_\theta\left(S_1(l_{i1}), S_2(u_{i2}), S_3(u_{i3})\right) - C_\theta\left(S_1(u_{i1}), S_2(u_{i2}), S_3(u_{i3})\right). \end{aligned}$$

The contribution of a cluster with two interval-censored observations (observation one and two) and one right-censored observation (observation three) is

$$\begin{aligned} L_{i,(1,1,0)} &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, l_{i2} \leqslant T_{i2} \leqslant u_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &\quad -P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right). \end{aligned} \tag{4.5}$$

The first term in the right hand side of (4.5) is equal to

$$P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right).$$

The second term in the right hand side of (4.5) is equal to

$$P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right) - P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant u_{i2}, T_{i3} \geqslant l_{i3}\right).$$

Bringing all the terms together, the contribution for a cluster with two interval-censored observations (observation 1 and 2) and one right-censored observation (observation 3) is

$$\begin{aligned} L_{i,(1,1,0)} &= S(l_{i1}, l_{i2}, l_{i3}) - S(u_{i1}, l_{i2}, l_{i3}) - S(l_{i1}, u_{i2}, l_{i3}) + S(u_{i1}, u_{i2}, l_{i3}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2}), S_3(l_{i3})\right) - C_\theta\left(S_1(u_{i1}), S_2(l_{i2}), S_3(l_{i3})\right) \\ &\quad -C_\theta\left(S_1(l_{i1}), S_2(u_{i2}), S_3(l_{i3})\right) + C_\theta\left(S_1(u_{i1}), S_2(u_{i2}), S_3(l_{i3})\right). \end{aligned}$$

The contributions where the first or second observation is right-censored and the other observations are interval-censored can be derived in a similar way. The contribution of a cluster with one interval-censored observation (observation one) and two right-censored observations (observations two and three) is

$$\begin{aligned} L_{i,(1,0,0)} &= P\left(l_{i1} \leqslant T_{i1} \leqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &\quad -P\left(T_{i1} \geqslant u_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &= S(l_{i1}, l_{i2}, l_{i3})) - S(u_{i1}, l_{i2}, l_{i3}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2}), S_3(l_{i3})\right) - C_\theta\left(S_1(u_{i1}), S_2(l_{i2}), S_3(l_{i3})\right). \end{aligned}$$

The contributions where the second or third observation is interval-censored and the other observations are right-censored can be derived in a similar way. A cluster with three right-censored observations has contribution

$$\begin{aligned} L_{i,(0,0,0)} &= P\left(T_{i1} \geqslant l_{i1}, T_{i2} \geqslant l_{i2}, T_{i3} \geqslant l_{i3}\right) \\ &= S(l_{i1}, l_{i2}, l_{i3}) \\ &= C_\theta\left(S_1(l_{i1}), S_2(l_{i2}), S_3(l_{i3})\right). \end{aligned}$$

In the contributions above $S(q, r, s)$ is the joint marginal survival function, the lower bound of the interval is used as the censoring time for a right-censored observation. In the copula model the joint marginal survival function is a copula function $C_\theta(v_1, v_2, v_3)$ with arguments (v_1, v_2, v_3) the marginal survival functions $S_1(q)$, $S_2(r)$ and $S_3(s)$. The loglikelihood consists of eight terms, representing the eight possible censoring configurations for trivariate data.

Fourdimensional data

The same reasoning can be applied in four dimensions. The likelihood for right- and interval-censored data consists of 16 contributions. Four contributions represent the case where one of the observations in a cluster is interval-censored and the others are right-censored ($L_{i,(1,0,0,0)}$, $L_{i,(0,1,0,0)}$, $L_{i,(0,0,1,0)}$ and $L_{i,(0,0,0,1)}$), six contributions represent the case where two of the observations are interval-censored and the others are right-censored ($L_{i,(1,1,0,0)}$, $L_{i,(1,0,1,0)}$, $L_{i,(1,0,0,1)}$, $L_{i,(0,1,1,0)}$, $L_{i,(0,1,0,1)}$ and $L_{i,(0,0,1,1)}$), four contributions represent the case where one of the observations is right-censored and the others are interval-censored ($L_{i,(1,1,1,0)}$, $L_{i,(1,1,0,1)}$, $L_{i,(1,0,1,1)}$ and $L_{i,(0,1,1,1)}$), one contribution stands for four right-censored observations ($L_{i,(0,0,0,0)}$) and one contribution depicts four interval-censored observations ($L_{i,(1,1,1,1)}$). The contributions are given by

$$\begin{aligned} &L_{i,(1,1,1,1)} = P\left(l_{i1} \leqslant T_1 \leqslant u_{i1}, l_{i2} \leqslant T_2 \leqslant u_{i2}, l_{i3} \leqslant T_3 \leqslant u_{i3}, l_{i4} \leqslant T_4 \leqslant u_{i4}\right) \\ &= S(l_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(u_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(l_{i1}, u_{i2}, l_{i3}, l_{i4}) + S(u_{i1}, u_{i2}, l_{i3}, l_{i4}) \\ &-S(l_{i1}, l_{i2}, u_{i3}, l_{i4}) + S(u_{i1}, l_{i2}, u_{i3}, l_{i4}) + S(l_{i1}, u_{i2}, u_{i3}, l_{i4}) - S(u_{i1}, u_{i2}, u_{i3}, l_{i4}) \\ &-S(l_{i1}, l_{i2}, l_{i3}, u_{i4}) + S(u_{i1}, l_{i2}, l_{i3}, u_{i4}) + S(l_{i1}, u_{i2}, l_{i3}, u_{i4}) - S(u_{i1}, u_{i2}, l_{i3}, u_{i4}) \\ &+S(l_{i1}, l_{i2}, u_{i3}, u_{i4}) - S(u_{i1}, l_{i2}, u_{i3}, u_{i4}) - S(l_{i1}, u_{i2}, u_{i3}, u_{i4}) + S(u_{i1}, u_{i2}, u_{i3}, u_{i4}) \end{aligned}$$

for a cluster with four interval-censored observations.

$$L_{i,(1,1,1,0)} = P\left(l_{i1} \leqslant T_1 \leqslant u_{i1}, l_{i2} \leqslant T_2 \leqslant u_{i2}, l_{i3} \leqslant T_3 \leqslant u_{i3}, u_{i4} \leqslant T_4\right)$$

$$= S(l_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(u_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(l_{i1}, u_{i2}, l_{i3}, l_{i4}) + S(u_{i1}, u_{i2}, l_{i3}, l_{i4})$$

$$-S(l_{i1}, l_{i2}, u_{i3}, l_{i4}) + S(u_{i1}, l_{i2}, u_{i3}, l_{i4}) + S(l_{i1}, u_{i2}, u_{i3}, l_{i4}) - S(u_{i1}, u_{i2}, u_{i3}, l_{i4})$$

for a cluster where the first three observations are interval-censored and the fourth observation is right-censored. The contributions for the other scenarios where one observation is right-censored can be obtained in a similar way.

$$L_{i,(1,1,0,0)} = P\left(l_{i1} \leqslant T_1 \leqslant u_{i1}, l_{i2} \leqslant T_2 \leqslant u_{i2}, l_{i3} \leqslant T_3, l_{i4} \leqslant T_4\right)$$

$$= S(l_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(u_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(l_{i1}, u_{i2}, l_{i3}, l_{i4}) + S(u_{i1}, u_{i2}, l_{i3}, l_{i4})$$

for a cluster where the first two observations are interval-censored and the third and fourth observation is right-censored. The contributions for the other scenarios where two observations are interval-censored can be obtained in a similar way.

$$\begin{aligned} L_{i,(1,0,0,0)} &= P\left(l_{i1} \leqslant T_1 \leqslant u_{i1}, l_{i2} \leqslant T_2, l_{i3} \leqslant T_3, l_{i4} \leqslant T_4\right) \\ &= S(l_{i1}, l_{i2}, l_{i3}, l_{i4}) - S(u_{i1}, l_{i2}, l_{i3}, l_{i4}) \end{aligned}$$

for a cluster where the first observation is interval-censored and the other observations are right-censored. The contributions for the other scenarios where one observation is interval-censored can be obtained in a similar way.

$$\begin{aligned} L_{i,(0,0,0,0)} &= P\left(l_{i1} \leqslant T_1, l_{i2} \leqslant T_2, l_{i3} \leqslant T_3, l_{i4} \leqslant T_4\right) \\ &= S(l_{i1}, l_{i2}, l_{i3}, l_{i4}). \end{aligned}$$

for a cluster with four right-censored observations. To write down the log-likelihood, we first introduce some additional notation, analogue to the no-

tation in Massonnet et al. (2009).

$$\begin{aligned}
\Delta_i &= \prod_{j=1}^{4}(1-\delta_{ij}) \\
\Delta_i(j) &= \delta_{ij}\prod_{k=1;k\neq j}^{4}(1-\delta_{ik}) \\
\Delta_i(j,k) &= \delta_{ij}\delta_{ik}\prod_{l=1;l\neq j,k}^{4}(1-\delta_{il}), \qquad j\neq k \\
\Delta_i(j,k,l) &= \delta_{ij}\delta_{ik}\delta_{il}(1-\delta_{im}), \qquad m\neq j,k,l; j\neq k, j\neq l, k\neq l \\
\Delta_i(1,2,3,4) &= \prod_{j=1}^{4}\delta_{ij}, \qquad (4.6)
\end{aligned}$$

with δ_{ij} equal to 1 if the j^{th} observation of the i^{th} cluster is interval-censored and equal to 0 if the j^{th} observation of the i^{th} cluster is right-censored. δ_{ik}, δ_{il}, δ_{im} are defined similarly. The joint marginal survival functions $S(p,q,r,s)$ in the contributions above can now be replaced by a copula function $C_\theta(v_1,v_2,v_3,v_4)$ with arguments (v_1,v_2,v_3,v_4) the marginal survival functions $S_1(p)$, $S_2(q)$, $S_3(r)$ and $S_4(s)$. The lower bound of the interval is used as the censoring time for a right-censored observation. The loglikelihood is then given by

$$\begin{aligned}
\log L(\boldsymbol{\zeta}) = \sum_{i=1}^{k}\Bigg[&\Delta_i \log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4}) \\
&+\sum_{j=1}^{4}\left[\Delta_i(j)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij})\right] \\
&+\sum_{j\neq k}\left[\Delta_i(j,k)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij},u_{ik})\right] \\
&+\sum_{j\neq k, j\neq l, k\neq l}\left[\Delta_i(j,k,l)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij},u_{ik},u_{il})\right] \\
&+\Delta_i(1,2,3,4)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{i1},u_{i2},u_{i3},u_{i4})\Bigg],
\end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi},\theta,\boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard, $\boldsymbol{\delta_i} = (\delta_{i1},\delta_{i2},\delta_{i3},\delta_{i4})$, and k the number of clusters. If a parametric distribution is chosen for the marginal survival functions, maximum likelihood

estimates can be obtained by maximizing this loglikelihood in a one-stage estimation approach using, for instance, the Newton Raphson procedure. Standard errors are obtained by taking the inverse of the Hessian matrix at the end of the optimization procedure.
In general, for a cluster with n members, there are 2^n contributions in the loglikelihood depending on the censoring status of the n members in the cluster. Moreover, the contribution representing n interval-censored observations also consists of 2^n terms. The contributions representing $n-l$ interval-censored observations, consist of 2^{n-l} terms, $l = 1, \ldots, n$. Therefore, the loglikelihood becomes complex with increasing cluster size and its practical use is limited to lower cluster sizes.

4.4.2 Analysis of the mastitis data

To analyze the mastitis data the following copula function for fourdimensional data is used

$$\begin{aligned} S(t_1,t_2,t_3,t_4) &= \Big[\{S_1(t_1)\}^{-\theta} + \{S_2(t_2)\}^{-\theta} \\ &\quad + \{S_3(t_3)\}^{-\theta} + \{S_4(t_4)\}^{-\theta} - 3\Big]^{-1/\theta}. \end{aligned}$$

A Weibull distribution is assumed for the marginal survival functions. The program is written in R. Parameter estimates and their standard errors obtained in the proposed model are given in Table 4.2. The rear udder quarters have a significantly higher hazard of infection than the front udder quarters for *Staph. aureus*, with HR = 1.32 (95% CI [1.05;1.66]), but a significantly lower hazard rate for the rear udder quarters is observed for *C. bovis*, with HR = 0.86 (95% CI [0.82;0.91]). For the two other bacteria, no significant differences were found, with the hazard ratio equal to 1.40 (95% CI [0.93;2.12]) for *Strep. dysgalactiae* and to 1.16 (95% CI [0.89;1.51]) for *Strep. uberis*. The hazard of infection for multiparous cows was significantly higher compared to heifers for *C. bovis* (HR = 1.41, 95% CI [1.23;1.61]) and *Strep. uberis* (HR = 2.18, 95% CI [1.37;3.45]). The hazard of infection for multiparous cows was also higher compared to heifers for *Staph. aureus* (HR = 1.24, 95% CI [0.87;1.76]) and *Strep. dysgalactiae* (HR = 1.11, 95% CI [0.67;1.82]), but not significantly.
The estimate for θ is 5.084 (0.830) for *Staph. aureus*, 3.244 (0.186) for *C. bovis*, 3.022 (1.266) for *Strep. dysgalactiae* and 5.033 (0.953) for *Strep. uberis*, with values of Kendall's τ equal to 0.72, 0.62, 0.60 and 0.72, respectively. The estimates obtained in the one-stage copula model are similar to the ones

obtained in the two-stage approach in which imputation of the midpoint was used in the second stage.

Table 4.2: Parameter estimates (Est) and their standard errors (SE) for the one-stage parametric copula model for fourdimensional interval-censored data with parity (with β_p the effect of a multiparous cow) and udder quarter location (with β_l the effect of a rear udder quarter) as covariates and Weibull baseline hazard for infection with either *Staphylococcus aureus*, *Corynebacterium bovis*, *Streptococcus dysgalactiae* or *Streptococcus uberis*.

Parameter	*Staphylococcus aureus* Est(SE)	*Corynebacterium bovis* Est(SE)	*Streptococcus dysgalactiae* Est(SE)	*Streptococcus uberis* Est(SE)
θ	5.084 (0.830)	3.244 (0.186)	3.022 (1.266)	5.033 (0.953)
λ	0.013 (0.002)	0.117 (0.008)	0.005 (0.002)	0.007 (0.002)
γ	1.009 (0.065)	1.261 (0.031)	1.005 (0.104)	1.039 (0.076)
β_l	0.278 (0.118)	-0.149 (0.027)	0.338 (0.210)	0.147 (0.135)
β_p	0.215 (0.180)	0.341 (0.068)	0.102 (0.253)	0.778 (0.235)

4.5 Conclusions

In this chapter we described some statistical techniques to model clustered, interval-censored data, available in commercial software packages. Only parametric models are considered, nonparametric and semiparametric models are much more complicated for interval-censored data. Some advantages and disadvantages of the different models are discussed. Especially in the fixed effects model caution is needed when introducing covariates at the cluster level because of the problem of confounding. It is important to interpret the hazard ratio correctly in the different models. In the marginal model and the copula model the hazard ratio represents the hazard of infection for a randomly chosen rear udder quarter versus the hazard of infection for a randomly chosen front udder quarter from whatever other cow. On the other hand, the fixed effects model is a conditional model. The hazard ratio reflects the hazard of infection of a rear versus a front udder quarter within the same cow. The copula model is the only model that provides an estimate for the clustering in the data. The different techniques were applied to the mastitis data, investigating the effect of covariates at the cow level and the

effect of covariates at the udder quarter level.
We also described an extension of the fourdimensional copula model for right-censored data to data that can be right-censored and interval-censored. In general the number of possible contributions in this likelihood is equal to 2^n with n the number of members in a cluster. Also, the number of terms in a contribution for a cluster with n interval-censored observations is equal to 2^n. Therefore, the likelihood becomes complex with increasing cluster size. It is recommended to restrict the use of the copula model for interval-censored data to lower dimensions. Application of this model to the mastitis data gives similar results as the copula model with midpoint imputation in the second stage.

Chapter 5

A shared gamma frailty model for multivariate interval-censored data

Based on:
Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P., and 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, 1–14.

5.1 Introduction

The most common and widely used survival analysis models are developed for independent, right-censored data. However, ordinary survival analysis techniques often need to be extended due to the particular data structure. The methodology presented in this chapter was especially developed for the mastitis data set, introduced in Section 1.9.2. This data set has two characteristics that require extension of the currently available survival analysis techniques if they have to be dealt with simultaneously. First, the data are hierarchically structured, with observation units (the udder quarters) grouped in clusters (the cow), so that the event times within a cow can not be assumed to be independent (Adkinson et al., 1993). Second, since the udder quarters are sampled only approximately monthly, the time to infection is not known exactly; it is only known that the infection happened between the last visit with a negative test and the first visit with a positive test, therefore, the infection time is interval-censored.

To handle the problem of interdependence for right-censored observations different models have been suggested among which the frailty model (see Section 1.7) is often used. For the analysis of independent interval-censored data a number of inferential techniques have been described (see Section 4.2). But analysis methods for settings where observations are at the same time correlated and interval-censored received less attention. Chapter 4 reviewed some methods available in commercial software packages and discussed their advantages and disadvantages. A fourdimensional one-stage copula model for interval-censored data was also described. Bellamy et al. (2004) proposed a method to fit clustered interval-censored data assuming a normal distribution for the random effect and integrating out the random effects numerically using Gaussian Quadrature. For more details on the model proposed by Bellamy et al. (2004) see Section 5.3.

In this chapter we propose an extension of the parametric shared gamma frailty model to interval-censored data. We show that a closed form expression of the marginal likelihood can be obtained by integrating out the gamma-distributed frailties, which can then be maximized to obtain parameter estimates. Variance estimates are obtained from the observed information matrix.

The technique allows the inclusion of covariates in the model. We want to investigate the effect of covariates that change within cow (e.g. front and rear udder quarters) and covariates that change between cows (e.g. parity, i.e., the number of previous calvings). But it also provides an estimate of the correlation between udder infection times within a cow, which is of in-

terest because it is a measure of the infectivity of the agent which causes the disease (Barkema et al., 1997).
The details on the model are given in Section 5.2. In Section 5.3 we illustrate the approach by analyzing the mastitis data. The performance of the method is evaluated based on a simulation study, presented in Section 5.4. Conclusions are given in the last section.

5.2 The parametric shared gamma frailty model with interval-censored data

We consider the following proportional hazard frailty model (see also Section 1.7.2)

$$h_{ij}(t) = h_0(t) z_i \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right), \quad i = 1, \ldots, k, j = 1, \ldots, n_i \tag{5.1}$$

with $h_{ij}(t)$ the hazard at time t for udder quarter j of cow i, $h_0(t)$ the baseline hazard at time t, $\boldsymbol{x}_{ij}$ the vector of covariates for the corresponding udder quarter and $\boldsymbol{\beta}$ the vector of covariate effects. We further assume that the frailties $z_1, \ldots, z_k$ are independent realizations from a one-parameter gamma distribution with mean one and variance θ (see (1.9)).
The udder quarter infection times in the mastitis study are either right-censored or interval-censored. Cluster i consists of $n_i = 4$ observations (one observation per udder quarter) of which r_i are right-censored and d_i are interval-censored. We write R_{ij} to denote the right-censored infection time for udder quarter j of cow i. If the information on the infection time is subject to interval censoring we denote the lower and upper bound of the interval as L_{ij} and U_{ij}. Per cluster we define two sets of indices according to whether the infection time is right-censored or interval-censored:

$$\begin{aligned} R_i &= \{j \in \{1,2,3,4\} : T_{ij} > R_{ij}\} \\ D_i &= \{j \in \{1,2,3,4\} : L_{ij} < T_{ij} \leqslant U_{ij}\}, \end{aligned}$$

with $R_i \cap D_i = \phi$ and $R_i \cup D_i = \{1,2,3,4\}$ and T_{ij} the unobservable infection time.
Assuming that the censoring process is not informative for the survival process (see Section 1.4.2) the conditional data likelihood contribution for cluster i consists of the product of differences of the conditional survival functions evaluated at the observed lower and upper time point for the interval-censored quarters and of the conditional survival function evaluated at the censoring time for the right-censored quarters

$$L_i\left(\theta, \boldsymbol{\xi}, \boldsymbol{\beta} \mid z_i\right) = \prod_{j \in R_i} S_{ij}(R_{ij}) \prod_{j \in D_i} \left\{S_{ij}(L_{ij}) - S_{ij}(U_{ij})\right\},$$

which results in

$$L_i\left(\theta, \boldsymbol{\xi}, \boldsymbol{\beta} \mid z_i\right) = \exp\left\{-\sum_{j \in R_i} H_{ij}(R_{ij})\right\} \times \prod_{j \in D_i} \left[\exp\left\{-H_{ij}(L_{ij})\right\} - \exp\left\{-H_{ij}(U_{ij})\right\}\right]$$

$$= \exp\left(-z_i C_i\right) \times \prod_{j \in D_i} \left\{\exp\left(-z_i L_{ij}^*\right) - \exp\left(-z_i U_{ij}^*\right)\right\}, \tag{5.2}$$

where $\boldsymbol{\xi}$ contains the parameters of the baseline hazard, $H_{ij}(.) = H_0(.) z_i \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)$, $C_i = \sum_{j \in R_i} H_0\left(R_{ij}\right) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)$, $L_{ij}^* = H_0\left(L_{ij}\right) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)$ and $U_{ij}^* = H_0\left(U_{ij}\right) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)$, with $H_0(.)$ the cumulative baseline hazard.
To be able to write down the product in the second factor of (5.2) in a general way, we define the following column vector $\boldsymbol{a}_i$ of length 2^{d_i} with d_i the number of elements in D_i

$$\boldsymbol{a}_i = \left({}_c a_{ik}\right)_{k=1}^{2^{d_i}} = \bigotimes_{j \in D_i} \begin{pmatrix} \exp(-z_i L_{ij}^*) \\ -\exp(-z_i U_{ij}^*) \end{pmatrix},$$

where $\bigotimes_{j \in D_i}$ represents the Kronecker product of the vectors $\left(\exp\left(-z_i L_{ij}^*\right), -\exp\left(-z_i U_{ij}^*\right)\right)^t, j \in D_i$. The first element of this column vector, for example, is $\exp\left(-z_i \sum_{j \in D_i} L_{ij}^*\right)$. The last element is $\pm \exp\left(-z_i \sum_{j \in D_i} U_{ij}^*\right)$ with a positive sign if the number of U_{ij}^*'s in the sum of the exponent is even and a negative sign if the number is odd. The number of U_{ij}^*'s in a_{ik} will be denoted as n_{ik}.
Expression (5.2) can then be rewritten as

$$L_i\left(\theta, \boldsymbol{\xi}, \boldsymbol{\beta} \mid z_i\right) = \exp\left(-z_i C_i\right) \left(\sum_{k=1}^{2^{d_i}} a_{ik}\right).$$

This expression still contains the unobserved frailty term z_i. We can however integrate out the frailty term which is assumed to have the gamma density

(1.9). We then obtain the marginal likelihood

$$\begin{aligned} L_i(\theta, \boldsymbol{\xi}, \boldsymbol{\beta}) &= \int_0^\infty \frac{z_i^{(1/\theta-1)} \exp(-z_i/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)} \exp(-z_i C_i) \left(\sum_{k=1}^{2^{d_i}} a_{ik} \right) dz_i \\ &= \frac{1}{\theta^{1/\theta}\Gamma(1/\theta)} \sum_{k=1}^{2^{d_i}} \int_0^\infty z_i^{(1/\theta-1)} \exp\left\{ -z_i \left(C_i + \frac{1}{\theta} \right) \right\} a_{ik} dz_i \\ &= \sum_{k=1}^{2^{d_i}} \frac{(-1)^{n_{ik}}}{\left(C_i + \frac{1}{\theta} + \log p_{ik} \right)^{1/\theta} \theta^{1/\theta}}, \end{aligned}$$

with $\boldsymbol{p}_i$ the column vector:

$$\boldsymbol{p}_i = ({}_c p_{ik})_{k=1}^{2^{d_i}} = \bigotimes_{j \in D_i} \begin{pmatrix} \exp(L_{ij}^*) \\ \exp(U_{ij}^*) \end{pmatrix}.$$

To obtain the full marginal likelihood (L) we take the product of the k cluster-specific marginal likelihoods $\prod_{i=1}^{k} L_i(\theta, \boldsymbol{\xi}, \boldsymbol{\beta})$. Maximum likelihood estimates can then be obtained by maximizing the full marginal likelihood using, for instance, the Newton Raphson procedure. As the second partial derivatives can be obtained for all parameters in the model (see the Appendix), an explicit expression for the information matrix is available, from which an estimate of the asymptotic variance-covariance matrix can be obtained. Different distributional assumptions for the baseline hazard are possible as will be discussed in the next section.

5.3 Analysis of the mastitis data

The proposed method will now be applied to the time to infection with *C. bovis* data set. 39.28% of the udder quarters were infected with *C. bovis* during the lactation period. We will investigate the effect of parity on the time to infection with *C. bovis*, which is a covariate at the cow level. Three categories will be considered: (i) primiparous cows (one calving, parity = 0), (ii) cows with between two and four calvings (parity = 1) and (iii) cows with more than four calvings (parity = 2). We have to categorize due to the fact that for some of the levels of parity only a small number of cows is available. We will also investigate whether there is a difference between front and rear udder quarters regarding to the time to infection with *C. bovis*. The location of the udder quarter (front or rear) is an udder quarter level covariate and thus changes within the cow.

Different choices of the distributional assumption for the baseline hazard are possible (Klein, 1992). Because of its mathematical simplicity we first look at the exponential distribution with constant hazard function $h_0(t) = \lambda$. As a constant hazard rate is probably not realistic in describing the time to intramammary *C. bovis* infection, we look at the Weibull distribution with hazard function $h_0(t) = \lambda\gamma t^{\gamma-1}$. The hazard is monotone decreasing for $\gamma < 1$ and monotone increasing for $\gamma > 1$ (see Section 1.5.1). We further consider another two-parameter distribution, the loglogistic distribution with hazard function $h_0(t) = \lambda\gamma t^{\gamma-1}/(1 + \lambda t^{\gamma})$. The numerator is the same as in the Weibull hazard, but the denominator makes that the hazard is monotone decreasing for $\gamma \leqslant 1$ and for $\gamma > 1$ the hazard increases initially to a maximum at time $\{(\gamma - 1)/\lambda\}^{1/\gamma}$ and then decreases to zero as time approaches infinity. The values of the loglikelihood are -6067.746 and -5651.909 for the model with an exponential and Weibull hazard, respectively; the Weibull distribution should definitely be preferred over the exponential (likelihood ratio test, p-value < 0.001). The comparison of the models with Weibull and loglogistic baseline hazards is based on the Akaike Information Criterion since these models are not nested within each other. To calculate the AIC we use the standard formula AIC $= -2 \log L + 2 \times$ (number of parameters) (Izumi and Ohtaki, 2004) . The AIC for the loglogistic and Weibull distribution are 11365.896 and 11315.818, respectively. Therefore, we will proceed with the Weibull distribution.

In order to assess the validity of the gamma frailty model with Weibull baseline hazard, we compare the marginal survival functions obtained by the nonparametric estimator for interval-censored data proposed by Turnbull (1976) (see Section 4.2.1) with the marginal survival function obtained from the gamma frailty model with Weibull baseline hazard. Figure 5.1 shows that for all possible combinations of the covariate levels for parity (category 0, 1 or 2) and udder quarter location (front or rear) the marginal survival function from the gamma frailty model follows the nonparametric estimate closely.

The parameter estimates obtained from the method proposed in the previous section with Weibull baseline hazard are shown in the first column of Table 5.1. Parameter estimates are obtained using infection times in terms of quarters of a year rather than days to avoid too small values for the estimate of λ. The parameter estimate $\hat{\lambda}$ referring to hazard rates in terms of quarters of a year can be back-transformed to the parameter corresponding to the daily hazard rate $\hat{\lambda}_d$ using $\lambda_d = \lambda \times (91.31)^{-\gamma}$. For the figures and the interpretation in the text, we make use of the rescale to days to make the interpretation easier.

Table 5.1: Parameter estimates (Est) and their standard errors (SE) for a gamma frailty model with parity (with $\hat{\beta}_{p_1}$ the effect of parity category 1 and $\hat{\beta}_{p_2}$ the effect of parity category 2) and udder quarter location (with $\hat{\beta}_l$ the effect of the rear udder quarter) as covariates and Weibull baseline hazard. Results are shown for the proposed method (exact), midpoint and upper bound of the interval as exact event times and Gaussian Quadrature (lognormal frailties).

	Exact Est (SE)	Midpoint Est (SE)	Upper bound Est (SE)	Gaussian Quadrature Est (SE)
θ	3.820 (0.223)	3.823 (0.222)	3.751 (0.219)	-
σ^2	-	-	-	4.949 (0.362)
λ	0.137 (0.016)	0.138 (0.016)	0.084 (0.010)	0.023 (0.003)
γ	1.987 (0.042)	1.984 (0.040)	2.437 (0.049)	2.092 (0.045)
β_l	-0.277 (0.050)	-0.276 (0.050)	-0.277 (0.050)	-0.279 (0.051)
β_{p1}	0.756 (0.148)	0.756 (0.147)	0.763 (0.148)	0.693 (0.169)
β_{p2}	1.309 (0.228)	1.307 (0.226)	1.334 (0.225)	1.244 (0.253)

The parameter estimate $\hat{\gamma}$ is above 1, and the hazard is thus increasing with time. The hazard ratio of cows with more than four calvings versus primiparous cows ($\exp(\hat{\beta}_{p2})$) equals 3.7 (p-value < 0.001) with 95% confidence interval [2.37;5.79]. The hazard ratio of cows with two to four calvings versus primiparous cows ($\exp(\hat{\beta}_{p1})$) equals 2.13 (p-value < 0.001) with 95% confidence interval [1.59;2.85]. The hazard ratio of rear versus front udder quarters ($\exp(\hat{\beta}_l)$) is 0.76 (p-value < 0.001) with 95% confidence interval [0.69;0.84]. The estimate for θ is 3.82 (0.223). From this estimate, Kendall's tau can be obtained and is equal to 0.66. Thus, infection times within the cow are highly correlated.

The proposed method is now compared to the naive method of imputing the midpoint or the upper bound of the interval as exact event time. Although using the midpoint gives us similar results (see Table 5.1), imputation of the upper bound has a large effect on the parameter estimates $\hat{\lambda}$ and $\hat{\gamma}$. Especially the overestimation of γ is eye-catching and leads to a more rapidly increasing hazard compared to the hazard obtained using imputation of midpoint or using the exact method based on interval-censored data. This can be seen clearly in Figure 5.2 in which the estimated hazard functions for the three models for a front udder quarter of a primiparous cow with frailty equal to one are depicted. The choice of the upper bound as exact event

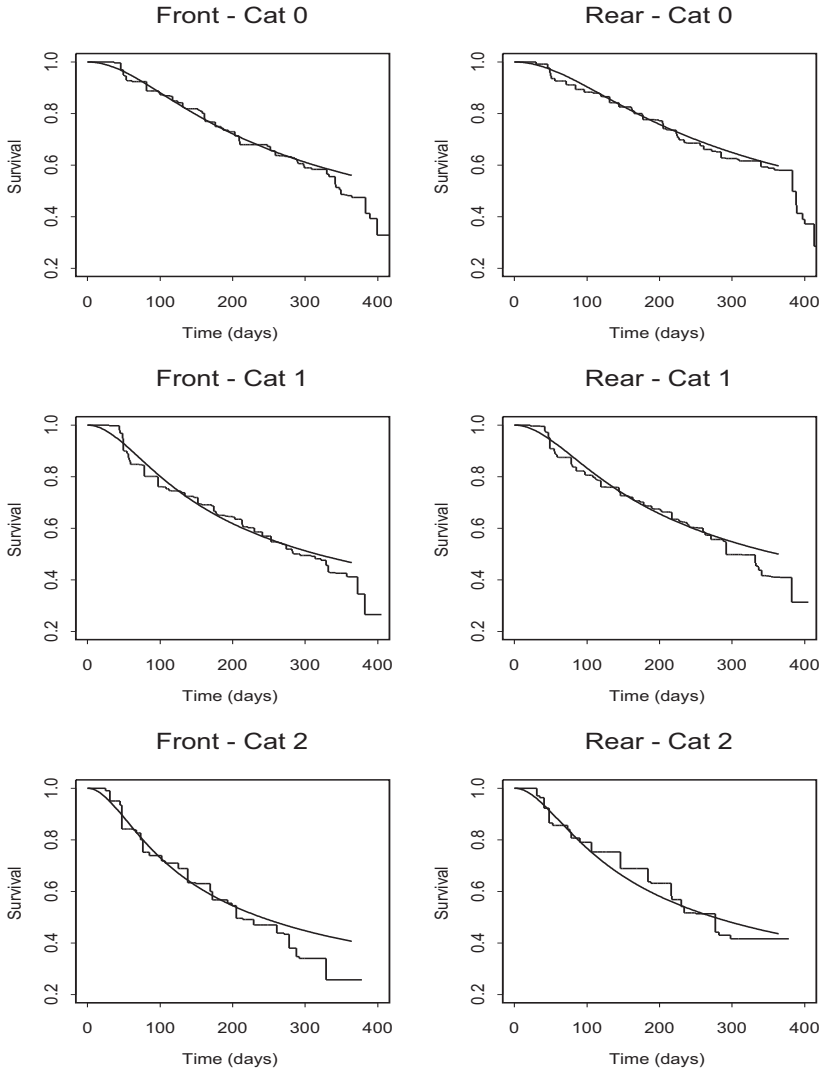

Figure 5.1: Estimated marginal survival functions for all possible combinations of the covariate levels of parity (category 0, 1 or 2) and udder quarter location (front or rear). The stepwise function corresponds to the nonparametric estimate for interval-censored data (Turnbull, 1976), whereas the continuous curve is obtained from the gamma frailty model with Weibull baseline hazard.

time makes that no events take place within the first 30 days after calving. Therefore, the model based on imputation of the upper bound leads to a faster increasing hazard function to accommodate for the fact that the hazard rate should be as low as possible in the first 30 days.

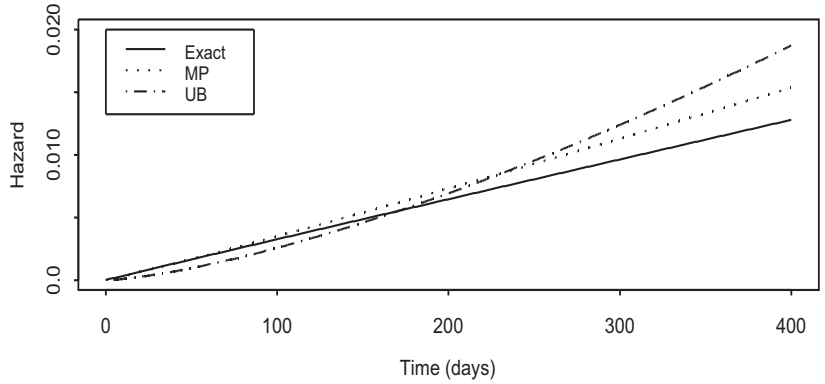

Figure 5.2: Estimated hazard functions from the gamma frailty model for a front udder quarter of a primiparous cow with frailty equal to one, either taking into account the interval censoring (Exact) or imputing the midpoint (MP) or the upper bound (UB).

It is also interesting to compare our results with the estimates obtained from the method proposed by Bellamy et al. (2004). In equation (5.1) the frailty z_i acts multiplicatively on the hazard rate and is gamma-distributed. The model formulation in Bellamy et al. (2004) expresses the frailty term as $\exp w_i$ with w_i the random effect working additively on the log hazard rate and assumes a normal distribution for the random effect (see Section 1.7.4). In what follows we write θ for the variance of the frailty z_i and σ^2 for the variance of the random effect w_i. We consider the following model

$$h_{ij}(t) = h_0(t) \exp(w_i) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right), \quad i = 1, \ldots, k, j = 1, \ldots, n_i \qquad (5.3)$$

with $h_0(t) = \lambda\gamma t^{\gamma-1}$ and $w_i \sim \mathrm{N}\left(0, \sigma^2\right)$. Since a normal distribution is assumed for the random effects w_i, the frailties follow a lognormal distribution and it is no longer possible to obtain a closed form expression for

the marginal likelihood by integrating out the frailties exactly. So Bellamy et al. (2004) used Gaussian Quadrature to integrate out the frailties and then maximized the marginal likelihood. Compared to Bellamy et al. (2004) we use the proportional hazards model representation (see Section 1.5.1) instead of the accelerated failure time model representation (see Section 1.5.2). With β_l the effect of the rear udder quarter, β_{p_1} the effect of the first parity category and β_{p_2} the effect of the second parity category, the parameter estimates correspond to $\hat{\sigma}^2 = 4.949(0.362)$, $\hat{\lambda} = 0.023(0.003)$, $\hat{\gamma} = 2.092(0.045)$, $\hat{\beta}_l = -0.279(0.051)$, $\hat{\beta}_{p_1} = 0.693(0.169)$ and $\hat{\beta}_{p_2} = 1.244(0.253)$.
It is not straightforward to compare the gamma frailty model (5.1) with model (5.3) with normally distributed random effects. The frailties corresponding to the random effects with mean equal to zero in the last model do not have mean one. In this particular case, the mean is estimated by $\exp\left(0.5\hat{\sigma}^2\right) = 11.88$. Therefore, it is good practice to compare the two models in terms of a medically relevant quantity such as the median time to infection (M). Both the random effects model and the frailty model induce heterogeneity in median time to infection between cows. The density function for the median time to infection in the gamma frailty model is given by

$$\begin{aligned} f_M(m) &= \gamma \left(\frac{\log 2}{\theta\lambda \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)} \right)^{1/\theta} \frac{1}{\Gamma\left(1/\theta\right)} \left(\frac{1}{m}\right)^{1+\frac{\gamma}{\theta}} \\ &\quad \exp\left(-\frac{\log 2}{\theta\lambda m^{\gamma} \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)} \right). \end{aligned}$$

In case of normally distributed random effects the density function for M is (Legrand et al., 2005)

$$f_M(m) = \frac{\gamma}{m\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left\{ \log\left(\frac{\log 2}{m\lambda \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)} \right) \right\}^2 \right].$$

The two density functions for a front udder quarter of a primiparous cow look rather similar (Figure 5.3), but compared to the model with normally distributed random effects, the gamma frailty model assumes that the median times to infection are less skew to the right and, therefore, have a somewhat higher peak at the more central median time to infection values. As a comparison to the results obtained in Chapter 4 the proposed method

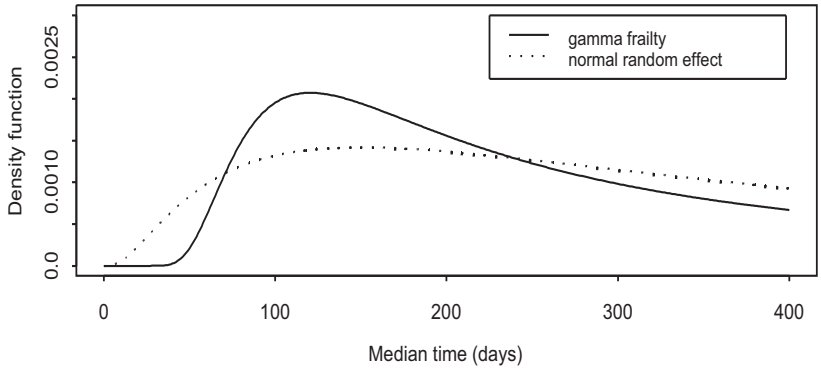

Figure 5.3: Density functions for the median time to infection from the frailty model with gamma and lognormal distributed frailty for the front udder quarter of a primiparous cow.

was also fit to the infection with *C. bovis*, *Staph. aureus*, *Strep. dysgalactiae* and *Strep. uberis* data sets with udder quarter location (front or rear) and dichotomous parity (multiparous or primiparous) as covariates. Results are given in Table 5.2.

The hazards of infection for the different bacteria are given in Figure 5.4 based on the frailty model and depicted for a cow with frailty equal to 1. The hazard of infection is about 10 times higher for *C. bovis* compared to the other bacteria. Furthermore, the hazard of infection continues to increase for *C. bovis* until the end of the lactation period, whereas for the other three bacteria, the hazard of infection levels off quickly to a constant level where it remains. The fact that the hazard of infection increases over time is reflected by the estimates of γ being all above 1. The rear udder quarters have a significantly higher hazard of infection than the front udder quarters for *Staph. aureus*, with HR = 1.40 (95% CI [1.08;1.83]), but a significantly lower hazard rate for the rear udder quarters is observed for *C. bovis*, with HR = 0.76 (95% CI [0.69;0.84]). For the two other bacteria, no significant differences were found, with the hazard ratio equal to 1.42 (95% CI [0.92;2.17]) for *Strep. dysgalactiae* and to 1.19 (95% CI [0.89;1.59]) for *Strep. uberis*. The hazard of infection for multiparous cows was significantly

Table 5.2: Parameter estimates (Est) and their standard errors (SE) for a gamma frailty model with parity (with $\hat{\beta}_p$ the effect of a multiparous cow) and udder quarter location (with $\hat{\beta}_l$ the effect of the rear udder quarter) as covariates and Weibull baseline hazard for infection with either *Staphylococcus aureus*, *Corynebacterium bovis*, *Streptococcus dysgalactiae* or *Streptococcus uberis*.

	Staphylococcus aureus Est (SE)	*Corynebacterium bovis* Est (SE)	*Streptococcus dysgalactiae* Est (SE)	*Streptococcus uberis* Est (SE)
θ	5.575 (0.921)	3.842 (0.224)	3.305 (1.392)	5.791 (1.077)
λ	0.014 (0.003)	0.137 (0.016)	0.005 (0.002)	0.007 (0.002)
γ	1.109 (0.070)	1.984 (0.042)	1.029 (0.106)	1.125 (0.080)
β_l	0.338 (0.135)	-0.276 (0.050)	0.349 (0.218)	0.175 (0.148)
β_p	0.238 (0.211)	0.867 (0.143)	0.110 (0.265)	0.888 (0.244)

higher compared to heifers for *C. bovis* (HR = 2.38, 95% CI [1.80;3.15]) and *Strep. uberis* (HR = 2.43, 95% CI [1.51;3.92]). The hazard of infection for multiparous cows was also higher compared to heifers for *Staph. aureus* (HR = 1.27, 95% CI [0.84;1.92]) and *Strep. dysgalactiae* (HR = 1.12, 95% CI [0.66;1.88]), but not significantly. Another important and interesting parameter of the frailty model is the variance of the frailties θ . The variance between cows is very high for all four bacteria. The highest estimate is obtained for *Strep. uberis* with $\hat{\theta} = 5.791(1.077)$, followed by *Staph. aureus* ($\hat{\theta} = 5.575(0.921)$), *C. bovis* ($\hat{\theta} = 3.842(0.224)$) and *Strep. dysgalactiae* ($\hat{\theta} = 3.305(1.392)$).

Though the parameter estimates are consequently lower in the marginal and copula models than in the frailty model for the *C. bovis* data set, the conclusions concerning the significance of the covariates for the different bacteria are the same in the frailty model and the marginal and copula models. The fixed effects model will not be included in our comparison since there are problems in estimating the covariate effects. The conclusions concerning the hazard are also similar especially for *Strep. dysgalactiae*, but some differences for *C. bovis*, *Staph. aureus* and *Strep. uberis* can be noted. The estimate for γ in the *C. bovis* data set is smaller in the marginal and copula models compared to the frailty model, leading to a less steeply increasing hazard. The estimate for γ in the *Staph. aureus* and *Strep.*

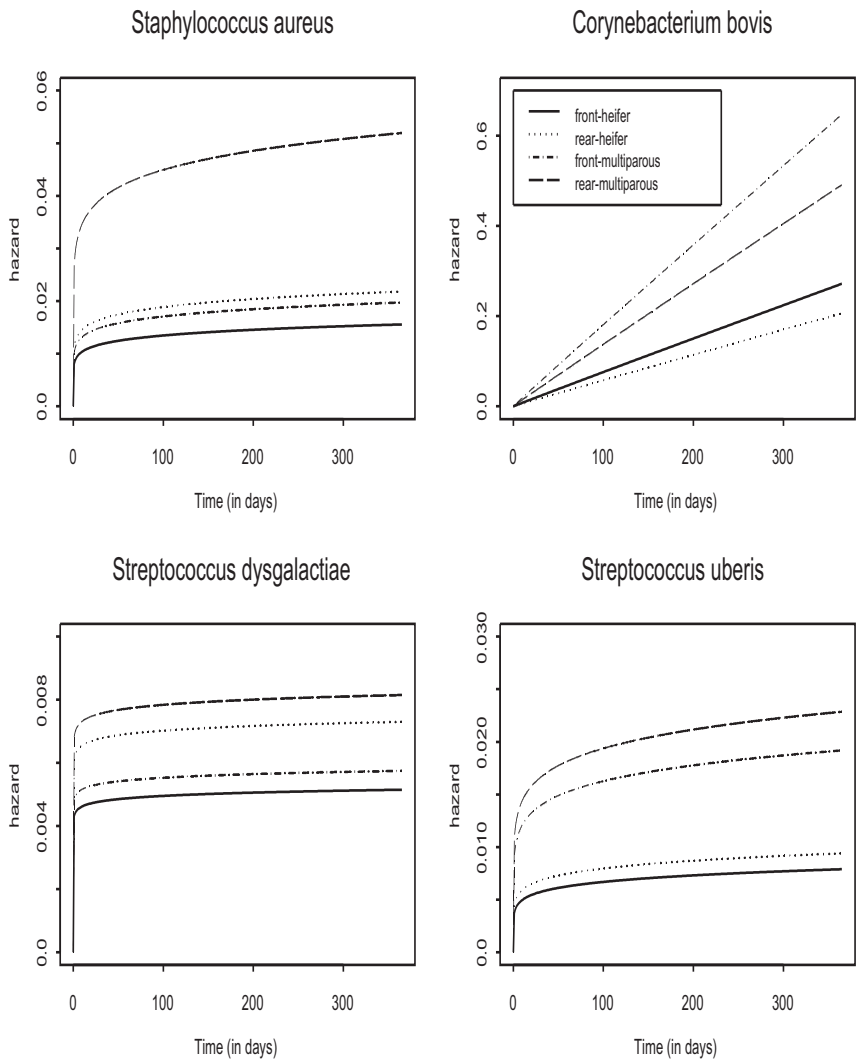

Figure 5.4: Estimated hazard functions for infection with either *Staphylococcus aureus*, *Corynebacterium bovis*, *Streptococcus dysgalactiae* or *Streptococcus uberis* based on the proposed frailty model for a cow with frailty equal to 1.

uberis data set is almost one in the marginal and copula models, leading to an almost constant hazard. In all three models θ is largest for infection with *Staph. aureus* and *Strep. uberis*.
Note that the introduction of the parity covariate has reduced the frailty variance estimate θ . For instance, for *Strep. uberis* the frailty variance estimate was equal to 6.322 without covariates. Thus the parity covariate can explain part of the between cows variability. This is not the case for the udder location covariate, as it is orthogonal to the frailty terms, meaning that it is independent of the frailty terms and will not alter the estimates of the frailty variance.

5.4 Simulation study

To evaluate the performance of the proposed methodology, a simulation study was done. For the simulation study, we first took a random subset of 100 cows of the mastitis data set, looking at infection with any bacterium. The true values for the simulation study then correspond to the parameter estimates obtained by fitting the proposed model to this subset with a single covariate, the udder location parameter. Using the same data structure as the example subset, 1000 data sets consisting of 100 clusters each, with four observations per cluster, were generated. The frailties (z_i) were generated from the one-parameter gamma distribution (1.9) with $\mathrm{E}(z_i) = 1$ and $\mathrm{Var}(z_i) = \theta = 1.8$. The data were simulated from the frailty model (5.1) assuming Weibull distributed event times with scale parameter λ equal to 0.9, shape parameter γ equal to 1.9 and β equal to 0.2. A binary covariate x takes the value 1 for the first two observations in a cluster and 0 for the other two. The percentage of censoring in the data sets lies around 25%, which is the percentage of censoring in the subset.
Asynchronous intervals of 30 days are generated around the simulated infection time as follows: the number of days a particular cow was in lactation before the first visit was simulated from the uniform distribution with a minimum of 1 and a maximum of 29. Visits are held at fixed time points (0-30-60-90-....) but since each cow entered the study at a different moment, namely the first day of its lactation, the endpoints of the intervals are adjusted to the number of days in lactation, so they can take any arbitrary value. All cows are assumed to be infection-free at the start of their lactation period, so if an udder quarter is already infected at the first visit, it is assumed that the infection took place between the start of the cow's lactation period and the first visit. The end of the study was set at one

year so that udder quarters with simulated infection time longer than one year are right-censored. The upper bound of the last interval is used as the censoring time.
For each of the 1000 data sets three models were fitted: the model proposed in Section 5.2 using the interval-censored data and the two naive models ignoring the interval censoring and imputing the midpoint or upper bound as exact event time.
The mean of the 1000 obtained estimates for the parameters θ, λ, γ and β is compared with the true value and differences between the three models are investigated. Standard errors are obtained by taking the inverse of the Hessian matrix at the end of the optimization procedure. The empirical standard error obtained from the 1000 data sets is also calculated. Finally, we also determine the coverage, defined as the percentage of the 1000 data sets that contains the true population parameter within their 95 % confidence interval.
The results of the simulation study suggest that the estimates obtained with our proposed model and by imputation of the midpoint are close to the true population parameters of interest (see Table 5.3). For the upper bound imputation, however, the estimate of λ is biased downward and the estimate of γ upward. The coverage is good if the interval-censored nature of the data is taken into account or if imputation of the midpoint is used. As can be expected because of the large bias for λ and γ, coverages for these parameters are unacceptable when the upper bound imputation is used. Based on these simulations, it might seem that our new technique has no advantage over imputation of the midpoint. However, this is not always the case. For instance, consider the same simulation setting as before ($\lambda = 0.9$, $\beta = 0.2$, $\theta = 1.8$, 30-day intervals) but change the value of the parameter γ to 0.5. Changing the value of the parameter γ from 1.9 to 0.5 means that the hazard is no longer increasing but decreasing over time. At the start of the study all udder quarters are at risk and, in case of an increasing hazard, few events take place in the beginning and a lot of udder quarters are still at risk towards the end of the study when more events take place. Therefore, a lot of information is available throughout the study and is used to obtain parameter estimates. For a decreasing hazard ($\gamma < 1$) a lot of events take place in the beginning leaving only few udder quarters at risk near the end of the study. So, when the interval-censored nature of the data is ignored in this setting, there is not enough information left to obtain adequate parameter estimates. As can be seen in Table 5.3 the exact method performs well in estimating all parameters including γ, but the techniques of imputing the midpoint or upper bound both fail in estimating the parameter γ and im-

Table 5.3: The averages of estimated model parameters (Est), the empirical standard error (Emp SE) and coverage (Cov) from 1000 simulated data sets using the proposed method (Exact), midpoint and upper bound of the interval as exact event times. True values for the parameters are given by $\lambda = 0.9$, $\beta = 0.2$, $\theta = 1.8$ and $\gamma = 1.9$ or 0.5 for the first resp. the second part of the table

	Exact Est (Emp SE, Cov)	Midpoint Est (Emp SE, Cov)	Upper bound Est (Emp SE, Cov)
		$\gamma = 1.9$	
$\hat{\lambda}$	0.896 (0.156, 93.6)	0.881 (0.151, 92.4)	0.596 (0.101, 21.5)
$\hat{\gamma}$	1.900 (0.104, 93.7)	1.887 (0.095, 93.9)	2.267 (0.115, 6.3)
$\hat{\beta}$	0.202 (0.124, 93.6)	0.198 (0.121, 94.1)	0.206 (0.126, 93.4)
$\hat{\theta}$	1.804 (0.285, 94.3)	1.760 (0.276, 93.1)	1.882 (0.286, 96.3)
		$\gamma = 0.5$	
$\hat{\lambda}$	0.917 (0.170, 94.5)	0.878 (0.181, 90.0)	0.709 (0.139, 62.1)
$\hat{\gamma}$	0.492 (0.039, 94.2)	0.735 (0.045, 0.0)	0.881 (0.056, 0.0)
$\hat{\beta}$	0.199 (0.145, 95.3)	0.212 (0.156, 93.4)	0.211 (0.156, 93.5)
$\hat{\theta}$	1.842 (0.359, 92.7)	2.141 (0.397, 89.7)	2.156 (0.398, 89.3)

putation also performs worse compared to the exact method in estimating the other parameters.

Situations in which censoring intervals are broader are also investigated. Therefore, intervals of width 60, 90 and 120 days are generated around the simulated infection times in the same manner as before. Figure 5.5 shows what happens with the estimates of the four model parameters when intervals become broader for the exact technique (empty box) and the midpoint imputation method (shaded box). The dashed horizontal line represents the true value of the parameter. The box represents the inter-quartile range and the solid line the median of the 1000 simulated data sets. The whiskers are drawn to the nearest value not beyond 1.5 times the inter-quartile range. It can be seen that the exact method performs better than the midpoint imputation approach in estimating the parameters θ, λ and β. The latter tends to underestimate these parameters. The bias is larger when intervals are broader. When the exact method is used, coverages are good, but

with the imputation technique coverages become smaller when intervals are broader. For example, when the interval spans 120 days, the coverage for λ is 59.8% with midpoint imputation versus 95.2% with the exact technique. The parameter γ however is underestimated by the exact technique while it is overestimated by midpoint imputation. For both techniques, coverages are smaller when intervals are broader. For the considered simulation studies it is clear that the exact method outperforms the two methods based on imputation.

5.5 Conclusions

In this chapter we proposed a shared gamma frailty model for clustered, interval-censored data with different baseline hazard functions. The model could even be made more flexible with the use of penalized splines for the baseline hazard using the same techniques (Rondeau et al., 2003). This is not discussed in this thesis. Assuming a gamma distribution for the frailty enables us to integrate out the frailties analytically and to obtain a closed form expression for the marginal likelihood, which can then be maximized with an optimization procedure such as the nlm function in R to obtain parameter estimates. Furthermore exact expressions for the second derivatives of the likelihood and thus estimates for the variances of the parameters can be obtained by inverting the matrix of second derivatives.
We compare our technique to the technique proposed by Bellamy et al. (2004) who propose normally distributed random effects. Under this assumption no closed form of the likelihood can be obtained and frailties are integrated out using Gaussian Quadrature. Some of the parameters appearing in the two models have the same meaning (γ and β) and can therefore be compared. To link the λ parameter in the proposed model to parameters in the model proposed by Bellamy et al. (2004) is more difficult. This is due to the specification of the cluster effects in terms of normally distributed random effects with mean zero, i.e. a lognormal distribution at the frailty level, but with a mean different from one. It is therefore not straightforward to compare these two hazard functions and we can expect large differences between them for large values of the variance of the lognormal and gamma distributions (Therneau and Grambsch, 2000). Indeed the larger the value of the variance parameter the more different the distributions are. The two models can however be meaningfully compared when the models are translated in terms of the density function of the median infection time. The two models result in comparable density functions.

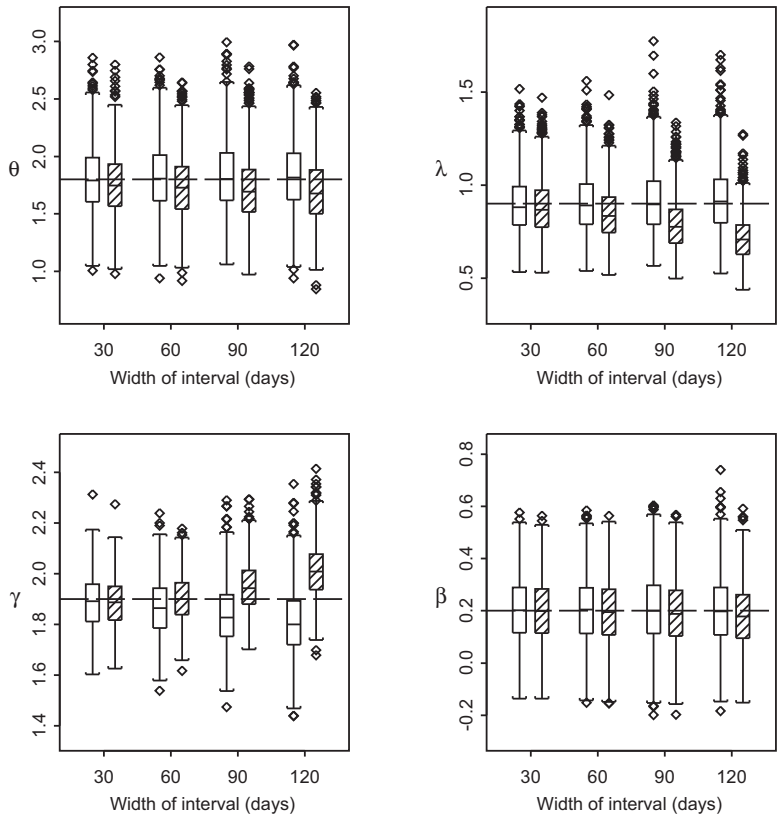

Figure 5.5: Boxplots for the estimated model parameters from 1000 simulated data sets using the proposed method (empty box) and the midpoint of the interval as exact event time (shaded box) for different interval widths (30-60-90-120). True values for the parameters are given by $\lambda = 0.9$, $\beta = 0.2$, $\theta = 1.8$ and $\gamma = 1.9$.

The proposed technique was applied to a data set consisting of 1196 cows who were approximately monthly screened for the presence of a bacterial infection at the udder quarter level, but the method is valid in a variety of situations since little or no data constraints apply to the proposed method. The data set needs to consist of a number of clusters from which the members can not be observed continuously. Contrary to the one-stage copula

model discussed in Chapter 4 the number of cluster members can be variable. Intervals can be of variable length, though the parameter γ tends to be more and more biased when intervals become broader. The first simulation setting shows that accurate estimates are obtained using the proposed technique and imputation of the midpoint. However, in the second simulation setting where the parameter γ is smaller than one, imputation of the midpoint fails. Using the upper bound as an exact event time leads to biased estimates especially for λ and γ in both simulation settings. Also, when censoring intervals get broader, using the midpoint imputation technique leads to biased estimates for all parameters. Overall, the simulation studies show that our technique outperforms the imputation techniques.

5.6 Appendix

5.6.1 Information matrix

The loglikelihood for cluster i in the proposed methodology is given by

$$l_i = \log = \sum_{k=1}^{2^{d_i}} \frac{(-1)^{n_{ik}}}{\left(C_i + \frac{1}{\theta} + \log p_{ik}\right)^{1/\theta} \theta^{1/\theta}}.$$

The second partial derivative for β in case of one covariate of the loglikelihood is given here as an example. The other partial derivatives can be obtained in a similar way.

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial \beta^2} &= (-1)\left\{\sum_{k=1}^{2^{d_i}} \frac{1}{\left(C_i + \log p_k + \frac{1}{\theta}\right)^{1/\theta}} \cdot s_k\right\} \\
&\quad -2 \cdot \left\{\frac{1}{\theta}\sum_{k=1}^{2^{d_i}} \frac{1}{\left(C_i + \log p_k + \frac{1}{\theta}\right)^{(1/\theta+1)}} \cdot (C_{i,\beta} + \log p_{k,\beta}) \cdot (-s_k)\right\}^2 \\
&\quad + \left\{\sum_{k=1}^{2^{d_i}} \frac{1}{\left(C_i + \log p_k + \frac{1}{\theta}\right)^{1/\theta}} \cdot s_k\right\}^{-1} \cdot \\
&\quad \left[\frac{1}{\theta}\sum_{k=1}^{2^{d_i}} \left\{\left(\frac{-1}{\theta} - 1\right) \frac{1}{\left(C_i + \log p_k + \frac{1}{\theta}\right)^{(1/\theta+2)}} \cdot (C_{i,\beta} + \log p_{k,\beta})^2\right.\right. \\
&\quad \left.\left. + \frac{1}{\left(C_i + \log p_k + \frac{1}{\theta}\right)^{(1/\theta+1)}} \cdot (C_{i,\beta\beta} + \log p_{k,\beta\beta})\right\}(-s_k)\right]
\end{aligned}$$

with $\mathbf{s}$ the following column vector:

$$\mathbf{s} = ({}_c s_k)_{k=1}^{2^{d_i}} = \bigotimes_{j=1}^{d_i} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and

$$C_{i,\beta} = \sum_{j \in R_i} H(R_{ij}) x_{ij} \exp(x_{ij}\beta)$$

$$\mathbf{p}_{k,\beta} = \bigotimes_{j=1}^{d_i} \begin{pmatrix} x_{ij} \exp(L_{ij}^*) \\ x_{ij} \exp(U_{ij}^*) \end{pmatrix}$$

$$C_{i,\beta\beta} = \sum_{j \in R_i} H(R_{ij}) (x_{ij})^2 \exp(x_{ij}\beta)$$

$$\mathbf{p}_{k,\beta\beta} = \bigotimes_{j=1}^{d_i} \begin{pmatrix} (x_{ij})^2 \exp(L_{ij}^*) \\ (x_{ij})^2 \exp(U_{ij}^*) \end{pmatrix}$$

To obtain the second partial derivative for the full marginal loglikelihood, we have to sum over the k cluster-specific second partial derivatives.

5.6.2 Software

Description of the variables

lower: the lower bound of the interval
upper: the upper bound of the interval
fail: the censoring indicator (1 in case of an event, 0 in case of right censoring)
X1, X2, X3: the covariates of interest
cluster: the cluster variable
theta, lambda, gamma and beta: the model parameters

The nlmixed-program

This program can be used to fit the model proposed by Bellamy et al. (2004) (see (5.3)) in SAS.

```
proc nlmixed data=Mastitis data qpoints=10 cov;
bounds gamma>0, theta>0, lambda>0;
G_t = exp(-exp(b)*lambda*(upper/91.31)**gamma
*exp(beta1*X1+beta2*X2+beta3*X3));
G1_t = exp(-exp(b)*lambda*(lower/91.31)**gamma
*exp(beta1*X1+beta2*X2+beta3*X3));
if fail=1 then lik=G1_t-G_t;
else if fail=0 then lik=G_t;
llik=log(lik);
model y~general(llik);
random b~normal(0,theta) subject =cluster;
run;
```

The R-program

This program can be used to fit the proposed methodology (see (5.1)) in R.

```
#Give the number of covariates
ncovar<-3

#Calculate the number of clusters.
clusternames<-levels(as.factor(cluster))
ncluster<-length(clusternames)

# Create a data set with the variables cluster,
# the lower bound,the upper bound,
# the censoring indicator and the covariates.
datasetint<-as.matrix(cbind(cluster,lower/91.31,
upper/91.31,fail,X1,X2,X3))

# create subsets for right-censored
and interval-censored observations
cendata<-datasetint[datasetint[,4]==0,]
intdata<-datasetint[datasetint[,4]==1,]

# Create a list of signs that corresponds to the n_ik
(here restricted to 4 events)
signs<-list(1,c(1,-1))
for(i in 3:5) signs[[i]]<-kronecker(signs[[i-1]],c(1,-1))

# Function to calculate the loglikelihood per cluster
```

```
CalcLogLikClust <-function(i,x)
{
theta<-exp(x[1])   lambda0<-exp(x[2]) gamma<-exp(x[3])
beta<-x[4:(3+ncovar)]
 if(ncovar==1) #univariate case
   {
cenX<-cendata[cendata[,1]==clusternames[i],5]
if (length(cenX)==0)Ci<-0
else {Ci<-sum(lambda*as.vector(cendata[cendata[,1]==
clusternames[i],3])^gamma*exp(cenX*beta))}
intL<-intdata[intdata[,1]==clusternames[i],2]

# if there are no events in that cluster
nevents <- length(intL)
crossprod <- 1
if(nevents>0)
{
intX<-intdata[intdata[,1]==clusternames[i],5]
intRster <- lambda*(intdata[intdata[,1]==
clusternames[i],3]^gamma)*exp(intX*beta)
intLster <- lambda*(intL^gamma)*exp(intX*beta)
crossprod<-c(exp(intLster[1]),exp(intRster[1]))
if(nevents>1)
{
for(ik in 2:nevents)
{
crossprod <- kronecker(crossprod,
c(exp(intLster[ik]),exp(intRster[ik])));
}
}
}
}
 else #multivariate
{
cenX<-cendata[cendata[,1]==clusternames[i],5:(4+ncovar)]
if (length(cendata[cendata[,1]==clusternames[i],5])==0)Ci<-0
else
{
Ci<-sum(lambda*(as.vector(cendata[cendata[,1]
==clusternames[i],3])^gamma)*exp(cenX%*%beta))
```

```
}

# if there are no events in that cluster
crossprod <- 1
intL<-intdata[intdata[,1]==clusternames[i],2]
nevents <- length(intL)
if(nevents>0)
{
intX<-intdata[intdata[,1]==clusternames[i],5:(4+ncovar)]
expiXb <- exp(intX%*%beta)
intRster <- lambda*(intdata[intdata[,1]
==clusternames[i],3]^gamma)
*expiXb
intLster <- lambda*(intL^gamma)*expiXb
crossprod<-c(exp(intLster[1]),exp(intRster[1]))
if(nevents>1)
{
for(ik in 2:nevents)
{
crossprod <- kronecker(crossprod,
c(exp(intLster[ik]),exp(intRster[ik])));
}
}
}

}

# Loglikelihood for 1 cluster
log(1/(theta^(1/theta))*sum((1/((sum(lambda*
(as.vector(cendata[cendata[,1]==clusternames[i],3])^gamma)
*exp(cenX*beta))+1/theta+log(crossprod))^(1/theta)))
*signs[[nevents+1]]))
}

# Calculate full marginal loglikelihood (formula 5)
CalcLogLik <- function(x)
{
-sum(sapply(1:ncluster,CalcLogLikClust,x=x))
}
```

```
# Maximising the full marginal loglikelihood
to obtain parameter estimates
init<-c(1,1,1,1)
print(results <- nlm(CalcLogLik,init,print.level=2,
 hessian=TRUE))

# Calculate covariance matrix
covmatr<-solve(results$hessian)
```

Chapter 6

The fourdimensional correlated gamma frailty model

Based on:
Goethals, K., Wienke, A., Janssen, P., and Duchateau, L. (2011), "Extensions of the correlated gamma frailty model to investigate the correlation structure between udder quarter infection times," *in preparation*.

6.1 Introduction

Shared frailty models have some limitations (Xue and Brookmeyer, 1996). First, the unobserved factors are necessarily the same for all members within a cluster, which is not always realistic. Second, shared frailty induces only positive correlation between the event times of the members in a cluster. However, in some cases event times for members within the same cluster may be negatively correlated. Third, Xue and Brookmeyer (1996) mention that the correlation between event times within a cluster is based on marginal distributions of event times. They refer to Clayton and Cuzick (1985) who discuss the confounding between the correlation parameter and the population heterogeneity when covariates are present in a proportional hazards model with gamma distributed frailty, implying that the joint distribution can be identified from the marginal distributions (Hougaard, 1986b). To avoid these problems, the correlated gamma frailty model was proposed by Yashin et al. (1995). It is a natural extension of the shared frailty model and of the univariate frailty model. In this model individual frailty consists of two parts: a part representing unobserved risk factors common for all members of a cluster, introducing correlation between the event times and a part representing individual unobserved risk factors. The combination of the two sources of heterogeneity forms the total individual frailty, therefore, the frailties of the members of a cluster are correlated but not necessarily shared (with correlation equal to one).
The remainder of this chapter is organized as follows: Section 6.2 describes the bivariate correlated gamma frailty model, in Section 6.3 different fourdimensional correlated gamma frailty models are introduced to model different correlation structures between the frailties of the four udder quarters in the mastitis data set. In Section 6.2 and Section 6.3 techniques are described for right-censored data; the midpoint of the interval is used as an exact event time. In Section 6.5 the interval-censored nature of the data is taken into account.

6.2 The bivariate correlated gamma frailty model

In the bivariate correlated frailty model frailties are not necessarily the same for the two subjects in a cluster. The bivariate correlated gamma frailty model was first introduced by Yashin et al. (1995) in the context of twin data. They used data on monozygotic and dizygotic twins to distinguish

genetic vs. environmental influences in the ageing process.
As in the univariate and shared frailty model, the frailty works multiplicatively on the baseline hazard and members of a pair are independent given the frailty. The correlated frailty model is given by

$$h_{ij}(t) = h_0(t) z_{ij} \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}),$$

with $h_{ij}(t)$ the conditional hazard function at time t for the j^{th} observation of the i^{th} pair, $j = 1, 2, i = 1, \ldots k$, $h_0(t)$ the baseline hazard, $\boldsymbol{x}_{ij}$ the vector of covariates for the j^{th} observation of the i^{th} pair, $\boldsymbol{\beta}$ the vector of regression parameters and z_{ij} the frailty for the j^{th} member of the i^{th} pair, coming from a density $f_Z(z)$. The correlation between the frailties of the members of a pair is denoted by ρ.
We assume in this chapter that the frailties follow a gamma distribution (1.9). The bivariate distribution of frailty is constructed using three independent gamma-distributed random variables. To ease notation we describe the construction of the bivariate distribution of frailty in one cluster, therefore dropping the index i. Assume that Y_0, Y_1 and Y_2 are independent gamma-distributed random variables with parameters (k_l, α_l), $l = 0, 1, 2$, respectively. Then assume that the frailties (Z_1, Z_2) for the two members of a pair are given by $Z_1 = Y_0 + Y_1$ and $Z_2 = Y_0 + Y_2$. Yashin et al. (1995) made two extra assumptions: they assumed that the scale parameters α_0, α_1 and α_2 of the gamma-distributed random variables Y_0, Y_1 and Y_2 are the same, i.e., $\alpha_0 = \alpha_1 = \alpha_2 = \alpha$. Furthermore they assumed that the shape parameters k_1 and k_2 for the distributions of Y_1 and Y_2 are the same, $k_1 = k_2$, forcing Z_1 and Z_2 to have the same distribution. This condition was relevant in their setting (twin studies), because there is no need for different distributions for the frailty for twin members, but can be omitted in other applications. Z_1 and Z_2 are then gamma-distributed random variables with parameters $(k_0 + k_1, \alpha)$. They are correlated since both contain the common part Y_0. We further add the restriction $k_0 + k_1 = \alpha$. The frailties Z_1 and Z_2 therefore have mean 1 and their variance is denoted by σ^2:

$$\begin{aligned} \mathrm{E}(Z_1) &= \mathrm{E}(Z_2) = \frac{k_0 + k_1}{\alpha} = 1 \\ \mathrm{V}(Z_1) &= \mathrm{V}(Z_2) = \frac{k_0 + k_1}{\alpha^2} = \frac{1}{\alpha} = \sigma^2. \end{aligned} \tag{6.1}$$

The correlation ρ between the frailties Z_1 and Z_2 can be calculated as

$$\rho = \mathrm{Corr}(Z_1, Z_2) = \frac{\mathrm{Cov}(Z_1, Z_2)}{\sqrt{\mathrm{V}(Z_1)\mathrm{V}(Z_2)}}.$$

The variance of Z_1 and Z_2 is given in (6.1), the covariance between Z_1 and Z_2 is calculated using the following formula

$$\text{Cov}(Z_1, Z_2) = \text{Cov}(Y_0 + Y_1, Y_0 + Y_2) = V(Y_0) = \frac{k_0}{\alpha^2}.$$

The covariance and correlation between Z_1 and Z_2 are therefore given by:

$$\begin{aligned} \text{Cov}(Z_1, Z_2) &= \frac{k_0}{\alpha^2} \\ \rho &= \frac{k_0}{\alpha}. \end{aligned}$$

Since $\rho = k_0/\alpha$, $k_0 = \rho\alpha = \rho/\sigma^2$ and since $(k_0 + k_1)/\alpha^2 = \sigma^2$, $k_1 = 1/\sigma^2 - k_0 = (1-\rho)/\sigma^2$. Using $k_0 \geqslant 0$ and $k_1 \geqslant 0$, it follows that

$$0 \leqslant \rho \leqslant 1.$$

It is further assumed that given the frailties Z_1 and Z_2, the event times T_1 and T_2 are independent. Under this condition the marginal bivariate survival function $S(t_1, t_2)$ can be obtained from the conditional bivariate survival function as follows:

$$\begin{aligned} S(t_1, t_2) &= \text{E}(S(t_1, t_2|Z_1, Z_2)) \\ &= \text{E}\left(S_1(t_1|Z_1)S_2(t_2|Z_2)\right) \\ &= \text{E}\left(e^{-Z_1 H_1(t_1)} e^{-Z_2 H_2(t_2)}\right) \\ &= \text{E}\left(e^{-(Y_0+Y_1)H_1(t_1)} e^{-(Y_0+Y_2)H_2(t_2)}\right) \\ &= \text{E}\left(e^{-Y_0(H_1(t_1)+H_2(t_2)) - Y_1 H_1(t_1) - Y_2 H_2(t_2)}\right). \end{aligned} \tag{6.2}$$

Making use of the Laplace transform of the gamma distribution (1.10), the marginal bivariate survival function (6.2) can be written as

$$\begin{aligned} S(t_1, t_2) &= \left(1 + \frac{H_1(t_1) + H_2(t_2)}{\alpha}\right)^{-k_0} \\ &\quad \left(1 + \frac{H_1(t_1)}{\alpha}\right)^{-k_1} \left(1 + \frac{H_2(t_2)}{\alpha}\right)^{-k_1} \\ &= \left(1 + \sigma^2 H_1(t_1) + \sigma^2 H_2(t_2)\right)^{\frac{-\rho}{\sigma^2}} \\ &\quad \left(1 + \sigma^2 H_1(t_1)\right)^{\frac{\rho-1}{\sigma^2}} \left(1 + \sigma^2 H_2(t_2)\right)^{\frac{\rho-1}{\sigma^2}}. \end{aligned}$$

In the univariate frailty model the marginal univariate survival function can be obtained as

$$S(t) = \mathcal{L}\left(H(t)\right) = \left(1 + \sigma^2 H(t)\right)^{(-1/\sigma^2)},$$

with $H(t)$ the cumulative baseline hazard. The marginal univariate density function can then be obtained as $f(t) = dS(t)/dt$. The cumulative baseline hazard is then

$$H(t) = \frac{S(t)^{-\sigma^2} - 1}{\sigma^2}. \tag{6.3}$$

Using (6.3) the marginal bivariate survival function can be written as

$$S(t_1, t_2) = \frac{\left(S_1(t_1) S_2(t_2)\right)^{1-\rho}}{\left(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1\right)^{\rho/\sigma^2}}. \tag{6.4}$$

Based on this joint survival function the likelihood for right-censored data can be constructed:

$$\begin{aligned} L(\boldsymbol{\zeta}) \quad = \quad & \prod_{i=1}^{k} \left(f(y_{i1}, y_{i2})\right)^{\delta_{i1}\delta_{i2}} \left(-\frac{\partial S(y_{i1}, y_{i2})}{\partial y_{i1}}\right)^{\delta_{i1}(1-\delta_{i2})} \\ & \left(-\frac{\partial S(y_{i1}, y_{i2})}{\partial y_{i2}}\right)^{(1-\delta_{i1})\delta_{i2}} \left(S(y_{i1}, y_{i2})\right)^{(1-\delta_{i1})(1-\delta_{i2})}, \end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \sigma^2, \rho, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard, k the number of clusters and $\delta_{ij}, j = 1, 2$ equal to one in case of an event and equal to zero in case of a censored observation. A cluster with two censored subjects has contribution $S(y_{i1}, y_{i2})$, a cluster with two event times has contribution $f(y_{i1}, y_{i2})$, the contribution of a cluster with one event time and one censored observation is $-\frac{\partial S(y_{i1}, y_{i2})}{\partial y_{i1}}$ $(-\frac{\partial S(y_{i1}, y_{i2})}{\partial y_{i2}})$ if we observe an event time for the first (second) subject and a censored observation for the second (first) subject.
The partial derivatives of the joint survival function, needed to construct the likelihood, are

$$\begin{aligned} \frac{\partial S(t_1, t_2)}{\partial t_j} \quad = \quad & -S(t_1, t_2) S_j(t_j)^{-\sigma^2 - 1} f_j(t_j) \\ & \left(\rho \left(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1\right)^{-1}\right. \\ & \left. + (1 - \rho) S_j(t_j)^{\sigma^2}\right) \qquad j = 1, 2 \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} &= S(t_1,t_2)\left(S_1(t_1)S_2(t_2)\right)^{-\sigma^2-1} f_1(t_1)f_2(t_2) \\ &\quad \left[\left(\rho\left(S_1(t_1)^{-\sigma^2}+S_2(t_2)^{-\sigma^2}-1\right)^{-1}+(1-\rho)S_1(t_1)^{\sigma^2}\right)\right. \\ &\quad \left(\rho\left(S_1(t_1)^{-\sigma^2}+S_2(t_2)^{-\sigma^2}-1\right)^{-1}+(1-\rho)S_2(t_2)^{\sigma^2}\right) \\ &\quad \left.+\rho\sigma^2\left(S_1(t_1)^{-\sigma^2}+S_2(t_2)^{-\sigma^2}-1\right)^{-2}\right] = f(t_1,t_2).\end{aligned}$$

The loglikelihood then takes the form

$$\begin{aligned}\log L(\boldsymbol{\zeta}) &= \sum_{i=1}^{k}\Bigg[(1-\delta_{i1})(1-\delta_{i2})\log S(y_{i1},y_{i2}) + \\ &\quad \delta_{i1}(1-\delta_{i2})\log\frac{-\partial S(y_{i1},y_{i2})}{\partial y_{i1}} + (1-\delta_{i1})\delta_{i2}\log\frac{-\partial S(y_{i1},y_{i2})}{\partial y_{i2}} \\ &\quad +\delta_{i1}\delta_{i2}\log f(y_{i1},y_{i2})\Bigg]. \qquad (6.5)\end{aligned}$$

If a parametric distribution is chosen for the baseline hazard, this closed form expression of the loglikelihood can be maximized with respect to the unknown parameters using standard maximization procedures, such as the Newton Raphson procedure. Standard errors can be obtained from the inverse of the observed information matrix. Representation (6.4) of the joint survival function also allows separate estimation of the marginal survival functions and the correlation parameters in a two-stage approach. The marginal survival functions can be estimated parametrically, nonparametrically or semiparametrically in the first stage and the estimated marginal survival functions can then be plugged into the loglikelihood (6.5) to obtain estimates for the parameters ρ and σ^2. This approach is therefore semiparametric if a nonparametric or semiparametric estimator is used for the marginal survival functions in the first stage. Using this approach however, a correlated copula model is fitted instead of a correlated frailty model, because the marginal survival functions no longer depend on the frailty variance. Parameter estimates and their interpretation will not be the same. Iachine (1995) proposes an extended version of the EM-algorithm appropriate for the analysis of bivariate survival data using the correlated gamma frailty model.

Without the two extra restrictions ($\alpha_0 = \alpha_1 = \alpha_2 = \alpha$ and $k_1 = k_2$) Y_0, Y_1 and Y_2 are independently gamma-distributed random variables: $Y_0 \sim$

gamma(k_0, α_0), $Y_1 \sim$ gamma(k_1, α_1) and $Y_2 \sim$ gamma(k_2, α_2). The individual frailties are then constructed as follows:

$$\begin{aligned} Z_1 &= \frac{\alpha_0}{\alpha_1} Y_0 + Y_1 \sim \text{gamma}(k_0 + k_1, \alpha_1) \\ Z_2 &= \frac{\alpha_0}{\alpha_2} Y_0 + Y_2 \sim \text{gamma}(k_0 + k_2, \alpha_2). \end{aligned}$$

The restrictions $k_0 + k_1 = \alpha_1$ and $k_0 + k_2 = \alpha_2$ are added. Therefore, the frailties Z_1 and Z_2 have mean 1, $\mathrm{E}(Z_1) = \mathrm{E}(Z_2) = 1$ and the variances are $\mathrm{V}(Z_1) = \frac{1}{\alpha_1} = \sigma_1^2$, $\mathrm{V}(Z_2) = \frac{1}{\alpha_2} = \sigma_2^2$. Assuming different variances for the different members in a pair can be necessary in some data sets, take for example pairs that consist of a father and his adopted son. In this setting there is need for different frailty distributions for the members of the pair. The covariance and correlation are now given by

$$\begin{aligned} \mathrm{Cov}(Z_1, Z_2) &= \frac{k_0}{\alpha_1 \alpha_2} \\ \rho &= \frac{k_0}{\sqrt{\alpha_1 \alpha_2}} = k_0 \sigma_1 \sigma_2. \end{aligned}$$

Consequently $k_0 = \frac{\rho}{\sigma_1 \sigma_2}$ and since $k_0 + k_l = \alpha_l = \frac{1}{\sigma_l^2}$, $k_l = \frac{1}{\sigma_l^2} - k_0 = \frac{1}{\sigma_l^2} - \frac{\rho}{\sigma_1 \sigma_2}$ (l=1,2). Using $k_0 \geqslant 0$, $k_1 \geqslant 0$ and $k_2 \geqslant 0$, it can be seen that the range of the correlation coefficient ρ depends on the values of σ_1 and σ_2:

$$0 \leqslant \rho \leqslant \min\left\{\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right\}.$$

If $\sigma_1^2 \neq \sigma_2^2$, the correlation between the frailties is always smaller than one. This restriction on the range of ρ can be a serious limitation, especially when the values of σ_1^2 and σ_2^2 differ substantially.
The marginal bivariate survival function is

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\frac{\sigma_1 \rho}{\sigma_2}} S_2(t_2)^{1-\frac{\sigma_2 \rho}{\sigma_1}}}{\left(S_1(t_1)^{-\sigma_1^2} + S_2(t_2)^{-\sigma_2^2} - 1\right)^{\rho/(\sigma_1 \sigma_2)}}.$$

The partial derivatives of the marginal bivariate survival function, needed to construct the likelihood, are

$$\begin{aligned} \frac{\partial S(t_1, t_2)}{\partial t_j} &= -S(t_1, t_2) f_j(t_j) S_j(t_j)^{-\sigma_j^2 - 1} \\ &\quad \left[\frac{\sigma_j \rho}{\sigma_i}\left(S_1(t_1)^{-\sigma_1^2} + S_2(t_2)^{-\sigma_2^2} - 1\right)^{-1}\right. \\ &\quad \left. + (1 - \frac{\sigma_j \rho}{\sigma_i}) S_j(t_j)^{\sigma_j^2}\right] \qquad j = 1, 2; i = 1, 2; i \neq j \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2)}{\partial t_1 t_2} &= S(t_1,t_2)S_1(t_1)^{-\sigma_1^2-1}S_2(t_2)^{-\sigma_2^2-1}f_1(t_1)f_2(t_2)\\ &\quad\left[\left[\frac{\sigma_1\rho}{\sigma_2}\left(S_1(t_1)^{-\sigma_1^2}+S_2(t_2)^{-\sigma_2^2}-1\right)^{-1}\right.\right.\\ &\quad\left.+\left(1-\frac{\sigma_1\rho}{\sigma_2}\right)S_1(t_1)^{\sigma_1^2}\right]\\ &\quad\left[\frac{\sigma_2\rho}{\sigma_1}\left(S_1(t_1)^{-\sigma_1^2}+S_2(t_2)^{-\sigma_2^2}-1\right)^{-1}\right.\\ &\quad\left.+\left(1-\frac{\sigma_2\rho}{\sigma_1}\right)S_2(t_2)^{\sigma_2^2}\right]\\ &\quad\left.+\left[\rho\sigma_1\sigma_2\left(S_1(t_1)^{-\sigma_1^2}+S_2(t_2)^{-\sigma_2^2}-1\right)^{-2}\right]\right]\end{aligned}$$

and the loglikelihood takes the form of (6.5). Parameter estimates can be obtained after maximization of the loglikelihood.

Though the gamma distribution is the most widely used frailty distribution in the correlated frailty model (Yashin et al., 1995; Zdravkovic et al., 2004; Wienke et al., 2005b), also other distributions have been considered for the frailties in the literature. A correlated lognormal frailty model was introduced by Xue and Brookmeyer (1996) and applied to a data set on mental health in patients under psychiatric care. Other authors that use the lognormal distribution as a frailty distribution include Yau and McGilchrist (1997) and Ripatti and Palmgren (2000). The advantage of the correlated lognormal frailty model over the correlated gamma frailty model is its flexibility because it is not based on an additive composition of the frailties as the correlated gamma frailty model. It is also easily extendable to more dimensions. On the other hand, it is no longer possible to obtain a closed form expression for the marginal likelihood and numerical integration is required to obtain parameter estimates. The power variance function is suggested as a frailty distribution in the correlated frailty model by Yashin and Iachine (1999). As an extension of this model Wienke et al. (2010) propose a bivariate correlated frailty model with compound Poisson frailty. It allows for a non-susceptible fraction in the population overcoming the common assumption in survival analysis that all subjects are susceptible to the event under study.

Though the correlation between the frailties in the correlated gamma frailty model discussed above is necessarily positive, Yashin and Iachine (1999) give an alternative derivation of the model, which does not make use of the concept of frailty, and allows a negative correlation between the frailties.

This is interesting for different reasons. First, this model can be used in the analysis of correlated event times where both positive and negative correlation may occur. Second, if the correlation between the frailties is allowed to be negative, $\rho = 0$ becomes an internal point of the parameter space and the classical likelihood ratio test to test $H_0 : \rho = 0$ has an asymptotic χ^2-distribution.
The identifiability of the correlated gamma frailty model is discussed in Yashin and Iachine (1997). Whereas the univariate gamma frailty model without observed covariates is not identifiable if the underlying hazard function is not specified parametrically, the correlated gamma frailty model and the shared gamma frailty model is identifiable without such specification.

6.3 The fourdimensional correlated gamma frailty model

Most applications of the correlated frailty model concern bivariate data, in particular in the context of genetics (see for example Yashin and Iachine (1997); Zdravkovic et al. (2004); Wienke et al. (2005b)). Correlated frailty models applied to cluster sizes larger than two and/or with application in a different discipline are found less in the literature.
Giard et al. (2002) suggest a fourdimensional correlated gamma frailty model to describe the ageing process of a twin pair. The ageing process is simplified to a process consisting of the three states 'healthy', 'ill' and 'deceased'. Therefore, two event times are considered for each twin member: the time to disease and the time to death, resulting in four possibly correlated event times for each twin pair. The frailties then represent susceptibility to disease or susceptibility to death. The model is applied to data on prostate cancer in male Swedish twins. Wienke et al. (2002) consider a competing risk situation in twins where censoring can be informative and non-informative. For each twin member the time to a first and second cause of death is considered, resulting in four possibly correlated event times for each twin pair. The frailties represent susceptibility to the two causes of death for the two members of the twin pair. The model is applied to cause-specific mortality data, where focus is on mortality from coronary heart disease. Jonker and Boomsma (2010) propose a fourdimensional correlated gamma frailty model to estimate the degree of heredity, environmental effects and twin effects of the age at which people contact a social helper for the first time. The data set consists of clusters of size four, each cluster consists of a twin pair and two other siblings.

In this section we describe fourdimensional correlated gamma frailty models to investigate the correlation structure between the frailties of the four udder quarters of a dairy cow. $T_{ij}, i = 1, \ldots, k, j = 1, 2, 3, 4$, is the time to infection with *C. bovis* and the frailties Z_{ij}, $i = 1, \ldots, k$, $j = 1, 2, 3, 4$, represent individual susceptibility to get infected for the four udder quarters.

6.3.1 The fourdimensional correlated gamma frailty model with equal correlation between the frailties (model 2)

The extension of the bivariate correlated gamma frailty model to the multivariate case is straightforward (Yashin and Iachine, 1999), but the likelihood function becomes complex with increasing cluster size. For right-censored data the likelihood in the bivariate case consists of four terms (one term for each censoring scenario), in the trivariate case there are eight likelihood contributions depending on the censoring status of the members in a cluster and with cluster size four sixteen terms can contribute to the likelihood. In general there are 2^n contributions to the likelihood with n the number of members within a cluster.
Figure 6.1 visualizes the construction of the frailties and the pairwise correlation structure for the model described in this section, applied to the infection with *C. bovis* data set. To ease notation the index i representing the cluster is dropped. Z_1, Z_2, Z_3 and Z_4 represent the frailty for the front left (FL), front right (FR), rear left (RL) and rear right (RR) udder quarter, respectively. The individual frailties in this fourdimensional corre-

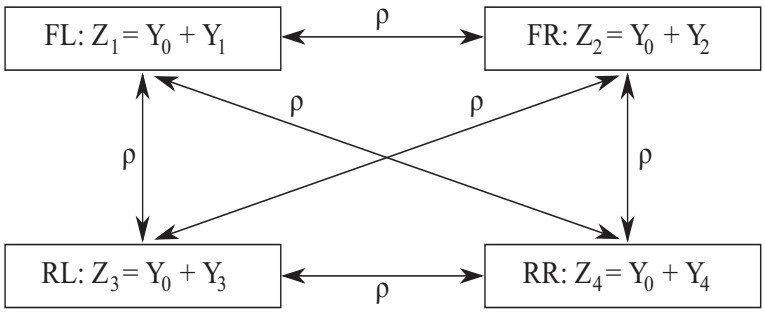

Figure 6.1: Construction of the frailties and the pairwise correlation structure for the fourdimensional correlated gamma frailty model with equal correlation (ρ) between the frailties. Z_1, Z_2, Z_3 and Z_4 represent the frailty for the front left (FL), front right (FR), rear left (RL) and rear right (RR) udder quarter, respectively.

lated gamma frailty model are constructed using five independent gamma-distributed random variables Y_0, Y_1, Y_2, Y_3 and Y_4 with parameters (k_l, α_l), l=0,. . . ,4, respectively. Y_0 represents the common part of the frailty and Y_1, Y_2, Y_3 and Y_4 represent the individual parts, of the frailty. We assume that the scale parameters of the gamma-distributed random variables Y_0, Y_1, Y_2, Y_3 and Y_4 are the same, i.e., $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha$. We further assume that the shape parameters for the distribution of Y_1, Y_2, Y_3 and Y_4 are the same, i.e., $k_1 = k_2 = k_3 = k_4$. The frailties are then given by

$$\begin{aligned} Z_1 &= Y_0 + Y_1 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_2 &= Y_0 + Y_2 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_3 &= Y_0 + Y_3 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_4 &= Y_0 + Y_4 \sim \text{gamma}(k_0 + k_1, \alpha). \end{aligned}$$

The restriction $k_0 + k_1 = \alpha$ is added. Therefore, the frailties have mean 1:

$$\text{E}(Z_1) = \text{E}(Z_2) = \text{E}(Z_3) = \text{E}(Z_4) = \frac{k_0 + k_1}{\alpha} = 1$$

and we denote their variance by σ^2

$$\text{V}(Z_1) = \text{V}(Z_2) = \text{V}(Z_3) = \text{V}(Z_4) = \frac{k_0 + k_1}{\alpha^2} = \frac{1}{\alpha} = \sigma^2.$$

The covariance between the frailties can be obtained as follows:

$$\text{Cov}(Z_i, Z_j) = \text{Cov}(Y_0 + Y_i, Y_0 + Y_j) = V(Y_0) = \frac{k_0}{\alpha^2}$$

with $i, j = 1 \ldots, 4; i < j$. The correlation between the frailties is then

$$\rho = \text{Corr}(Z_i, Z_j) = \frac{\text{Cov}(Z_i, Z_j)}{\sqrt{\text{V}(Z_i)\text{V}(Z_j)}} = \frac{k_0}{\alpha}.$$

Since $\rho = \frac{k_0}{\alpha}$, $k_0 = \rho\alpha = \frac{\rho}{\sigma^2}$ and since $\frac{k_0+k_1}{\alpha^2} = \sigma^2$, $k_1 = \frac{1}{\sigma^2} - k_0 = \frac{1-\rho}{\sigma^2}$. Using $k_0 \geqslant 0$ and $k_1 \geqslant 0$, the range of the correlation coefficient ρ is

$$0 \leqslant \rho \leqslant 1.$$

We further assume that given the frailties Z_1, Z_2, Z_3 and Z_4, the event times T_1, T_2, T_3 and T_4 are independent. Under this condition the marginal survival function $S(t_1, t_2, t_3, t_4)$ can be obtained from the conditional survival

function as follows:

$$\begin{aligned}
S(t_1,t_2,t_3,t_4) &= \mathrm{E}(S(t_1,t_2,t_3,t_4|Z_1,Z_2,Z_3,Z_4)) \\
&= \mathrm{E}\left(S_1(t_1|Z_1)S_2(t_2|Z_2)S_3(t_3|Z_3)S_4(t_4|Z_4)\right) \\
&= \mathrm{E}\left(e^{-Z_1H_1(t_1)}e^{-Z_2H_2(t_2)}e^{-Z_3H_3(t_3)}e^{-Z_4H_4(t_4)}\right) \\
&= \mathrm{E}\Big(e^{-(Y_0+Y_1)H_1(t_1)}e^{-(Y_0+Y_2)H_2(t_2)} \\
&\quad e^{-(Y_0+Y_3)H_3(t_3)}e^{-(Y_0+Y_4)H_4(t_4)}\Big) \\
&= \mathrm{E}\Big(e^{-Y_0(H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4))} \\
&\quad e^{-Y_1H_1(t_1)}e^{-Y_2H_2(t_2)}e^{-Y_3H_3(t_3)}e^{-Y_4H_4(t_4)}\Big). \qquad (6.6)
\end{aligned}$$

Making use of the Laplace transform of the gamma distribution (1.10), the marginal joint survival function (6.6) can be written as

$$\begin{aligned}
S(t_1,t_2,t_3,t_4) &= \left(1+\frac{H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4)}{\alpha}\right)^{-k_0} \\
&\quad \left(1+\frac{H_1(t_1)}{\alpha}\right)^{-k_1}\left(1+\frac{H_2(t_2)}{\alpha}\right)^{-k_1} \\
&\quad \left(1+\frac{H_3(t_3)}{\alpha}\right)^{-k_1}\left(1+\frac{H_4(t_4)}{\alpha}\right)^{-k_1} \\
&= \left(1+\sigma^2H_1(t_1)+\sigma^2H_2(t_2)+\sigma^2H_3(t_3)+\sigma^2H_4(t_4)\right)^{\frac{-\rho}{\sigma^2}} \\
&\quad \left(1+\sigma^2H_1(t_1)\right)^{\frac{\rho-1}{\sigma^2}}\left(1+\sigma^2H_2(t_2)\right)^{\frac{\rho-1}{\sigma^2}} \\
&\quad \left(1+\sigma^2H_3(t_3)\right)^{\frac{\rho-1}{\sigma^2}}\left(1+\sigma^2H_4(t_4)\right)^{\frac{\rho-1}{\sigma^2}}.
\end{aligned}$$

Using relationship (6.3) the marginal survival function can be written as

$$S(t_1,t_2,t_3,t_4) = \left(S_1(t_1)S_2(t_2)S_3(t_3)S_4(t_4)\right)^{1-\rho}\,\mathrm{A}^{-\rho/\sigma^2}, \qquad (6.7)$$

with $\mathrm{A} = S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} + S_3(t_3)^{-\sigma^2} + S_4(t_4)^{-\sigma^2} - 3$. For $\rho = 1$ the shared gamma frailty model is obtained, for $\rho = 0$ the univariate frailty model is obtained.
The partial derivatives of the marginal survival function, needed to construct the likelihood, are

$$\begin{aligned}
\frac{\partial S(t_1,t_2,t_3,t_4)}{\partial t_j} &= -S(t_1,t_2,t_3,t_4)S_j(t_j)^{-\sigma^2-1}f_j(t_j) \\
&\quad \left[\rho A^{-1}+(1-\rho)S_j(t_j)^{\sigma^2}\right] \qquad \text{for } j=1,2,3,4
\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} &= S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1} f_j(t_j) f_k(t_k) \\ &\quad \Big[\big[\rho A^{-1} + (1-\rho)S_k(t_k)^{\sigma^2}\big]\big[\rho A^{-1} + (1-\rho)S_j(t_j)^{\sigma^2}\big] \\ &\quad +\sigma^2 \rho A^{-2}\Big] \qquad \text{for } j,k = 1,2,3,4; j<k\end{aligned}$$

$$\begin{aligned}\frac{\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l} &= -S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k)S_l(t_l))^{-\sigma^2-1} \\ &\quad f_j(t_j) f_k(t_k) f_l(t_l) \\ &\quad \Big\{\big[\rho A^{-1} + (1-\rho)S_l(t_l)^{\sigma^2}\big]\Big[\big[\rho A^{-1} + (1-\rho)S_k(t_k)^{\sigma^2}\big] \\ &\quad \big[\rho A^{-1} + (1-\rho)S_j(t_j)^{\sigma^2}\big] + \sigma^2\rho A^{-2}\Big] \\ &\quad +\big[\rho A^{-1} + (1-\rho)S_j(t_j)^{\sigma^2}\big]\big[\sigma^2\rho A^{-2}\big] \\ &\quad +\big[\rho A^{-1} + (1-\rho)S_k(t_k)^{\sigma^2}\big]\big[\sigma^2\rho A^{-2}\big] \\ &\quad +2\sigma^4\rho A^{-3}\Big\} \qquad \text{for } j,k,l = 1,2,3,4; j<k<l\end{aligned}$$

$$\begin{aligned}\frac{\partial^4 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l \partial t_m} &= S(t_1,t_2,t_3,t_4)\Big(S_j(t_j)S_k(t_k)S_l(t_l)S_m(t_m)\Big)^{-\sigma^2-1} \\ &\quad f_j(t_j) f_k(t_k) f_l(t_l) f_m(t_m) \\ &\quad \Big\{\big[\rho A^{-1} + (1-\rho)S_m(t_m)^{\sigma^2}\big]\Big[\big[\rho A^{-1} + (1-\rho)S_l(t_l)^{\sigma^2}\big] \\ &\quad \Big[\big[\rho A^{-1} + (1-\rho)S_k(t_k)^{\sigma^2}\big]\big[\rho A^{-1} + (1-\rho)S_j(t_j)^{\sigma^2}\big] \\ &\quad +\sigma^2\rho A^{-2}\Big] + \big[\rho A^{-1} + (1-\rho)S_j(t_j)^{\sigma^2}\big]\big[\sigma^2\rho A^{-2}\big] \\ &\quad +\big[\rho A^{-1} + (1-\rho)S_k(t_k)^{\sigma^2}\big]\big[\sigma^2\rho A^{-2}\big] + 2\sigma^4\rho A^{-3}\Big]\end{aligned}$$

$$
\begin{aligned}
&+\left[\sigma^2\rho A^{-2}\right]\Big[\left[\rho A^{-1}+(1-\rho)S_k(t_k)^{\sigma^2}\right]\\
&\left[\rho A^{-1}+(1-\rho)S_j(t_j)^{\sigma^2}\right]+\sigma^2\rho A^{-2}\Big]\\
&+\left[\rho A^{-1}+(1-\rho)S_l(t_l)^{\sigma^2}\right]\Big[\left[\sigma^2\rho A^{-2}\right]\\
&\left[\rho A^{-1}+(1-\rho)S_j(t_j)^{\sigma^2}\right]\\
&+\left[\rho A^{-1}+(1-\rho)S_k(t_k)^{\sigma^2}\right]\left[\sigma^2\rho A^{-2}\right]+2\sigma^4\rho A^{-3}\Big]\\
&+\left[2\sigma^4\rho A^{-3}\right]\left[\rho A^{-1}+(1-\rho)S_j(t_j)^{\sigma^2}\right]\\
&+2\left[\sigma^2\rho A^{-2}\right]^2+\left[\rho A^{-1}+(1-\rho)S_k(t_k)^{\sigma^2}\right]\left[2\sigma^4\rho A^{-3}\right]\\
&+6\sigma^6\rho A^{-4}\Big\}. \quad \text{for } j=1;k=2;l=3;m=4
\end{aligned}
$$

To write down the loglikelihood, we first introduce some additional notation, analogue to the notation in Section 4.4.

$$
\begin{aligned}
\Delta_i &= \prod_{j=1}^{4}(1-\delta_{ij})\\
\Delta_i(j) &= \delta_{ij}\prod_{k=1;k\neq j}^{4}(1-\delta_{ik})\\
\Delta_i(j,k) &= \delta_{ij}\delta_{ik}\prod_{l=1;l\neq j,k}^{4}(1-\delta_{il}), \qquad j\neq k\\
\Delta_i(j,k,l) &= \delta_{ij}\delta_{ik}\delta_{il}(1-\delta_{im}), \qquad m\neq j,k,l; j\neq k; j\neq l, k\neq l\\
\Delta_i(1,2,3,4) &= \prod_{j=1}^{4}\delta_{ij}, \qquad (6.8)
\end{aligned}
$$

with δ_{ij} equal to 1 if the j^{th} member of the i^{th} cluster is interval-censored and equal to 0 if the j^{th} member of the i^{th} cluster is right-censored. The

loglikelihood is then given by

$$\begin{aligned}\log L(\boldsymbol{\zeta}) \;=\; & \sum_{i=1}^{k}\Bigg[\Delta_i \log S(t_1,t_2,t_3,t_4) \\ & + \sum_{j=1}^{4}\left[\Delta_i(j)\log\frac{-\partial S(t_1,t_2,t_3,t_4)}{\partial t_j}\right] \\ & + \sum_{j\neq k}\left[\Delta_i(j,k)\log\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k}\right] \\ & + \sum_{j\neq k;j\neq l;k\neq l}\left[\Delta_i(j,k,l)\log\frac{-\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l}\right] \\ & + \Delta_i(1,2,3,4)\log\frac{\partial^4 S(t_1,t_2,t_3,t_4)}{\partial t_1 \partial t_2 \partial t_3 \partial t_4}\Bigg]. \end{aligned} \tag{6.9}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \sigma^2, \rho, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard and k the number of clusters. If a parametric distribution is chosen for the baseline hazard, this closed form expression of the loglikelihood can be maximized with respect to the unknown parameters. Standard errors can be obtained from the inverse of the observed information matrix. Representation (6.7) of the joint survival function also allows a semiparametric two-stage approach. The marginal survival functions can be estimated nonparametrically or semiparametrically in the first stage and the estimated marginal survival functions can then be plugged into the loglikelihood (6.9) to obtain estimates for the parameters ρ and σ^2. In this approach a correlated copula model is fitted to the data instead of a correlated frailty model.

6.3.2 The fourdimensional correlated gamma frailty model with shared and correlated frailties (model 3)

We now present a fourdimensional correlated gamma frailty model where the correlation structure between the frailties is not symmetric, i.e., not all pairs of frailties have the same correlation. The correlation between the frailties of a pair can be different depending on the specific pair considered. It could make sense to assume an asymmetric correlation structure between the udder quarters of a dairy cow. Correlation between the two front udder quarters and the two rear udder quarters could be higher than between any other udder quarter pair. Similarly the two left udder quarters and the two

right udder quarters could be more correlated than any other udder quarter pair. Figure 6.2 visualizes the construction of the frailties and the pairwise correlation structure for the model described in this section. The individual

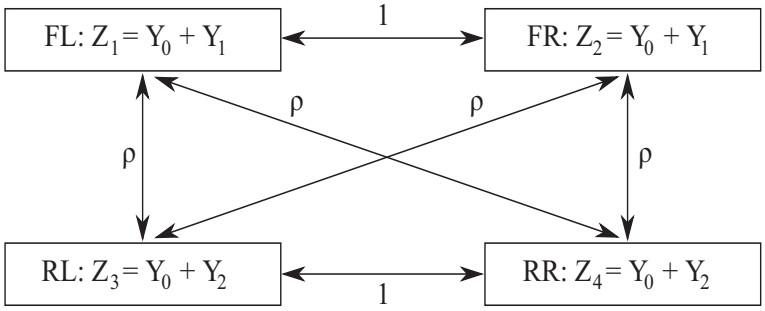

Figure 6.2: Construction of the frailties and the pairwise correlation structure for the fourdimensional correlated frailty model with correlations between the frailties equal to one or ρ. Z_1, Z_2, Z_3 and Z_4 represent the frailty for the front left (FL), front right (FR), rear left (RL) and rear right (RR) udder quarter, respectively.

frailties are constructed using three independent gamma-distributed random variables Y_0, Y_1 and Y_2 with parameters (k_l, α_l), l=0,1,2, respectively. Y_0 represents the common part of the frailty for each member of the cluster. Y_1 and Y_2 are common for only two members of the cluster, for example Y_1 is common for the first two members and Y_2 is common for the other two members. There is no variable representing truly individual heterogeneity. We again assume that the scale parameters of the gamma-distributed random variables Y_0, Y_1, Y_2 are the same, i.e., $\alpha_0 = \alpha_1 = \alpha_2 = \alpha$. We further assume that the shape parameters for the distribution of Y_1 and Y_2 are the same, i.e., $k_1 = k_2$. The frailties are then given by

$$\begin{aligned} Z_1 &= Y_0 + Y_1 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_2 &= Y_0 + Y_1 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_3 &= Y_0 + Y_2 \sim \text{gamma}(k_0 + k_1, \alpha) \\ Z_4 &= Y_0 + Y_2 \sim \text{gamma}(k_0 + k_1, \alpha) \end{aligned}$$

We further add the restriction $k_0 + k_1 = \alpha$. The frailties therefore have mean 1:

$$\text{E}(Z_1) = \text{E}(Z_2) = \text{E}(Z_3) = \text{E}(Z_4) = \frac{k_0 + k_1}{\alpha} = 1$$

and we denote their variance by σ^2

$$\mathrm{V}(Z_1) = \mathrm{V}(Z_2) = \mathrm{V}(Z_3) = \mathrm{V}(Z_4) = \frac{k_0 + k_1}{\alpha^2} = \frac{1}{\alpha} = \sigma^2.$$

The covariances between the frailties can be obtained as follows:

$$\begin{aligned}
\mathrm{Cov}(Z_1, Z_2) &= \mathrm{Cov}(Y_0 + Y_1, Y_0 + Y_1) = V(Y_0 + Y_1) = \frac{k_0 + k_1}{\alpha^2} \\
\mathrm{Cov}(Z_3, Z_4) &= \mathrm{Cov}(Y_0 + Y_2, Y_0 + Y_2) = V(Y_0 + Y_2) = \frac{k_0 + k_1}{\alpha^2} \\
\mathrm{Cov}(Z_1, Z_3) &= \mathrm{Cov}(Y_0 + Y_1, Y_0 + Y_2) = V(Y_0) = \frac{k_0}{\alpha^2} \\
&= \mathrm{Cov}(Z_1, Z_4) \\
&= \mathrm{Cov}(Z_2, Z_3) \\
&= \mathrm{Cov}(Z_2, Z_4).
\end{aligned}$$

The correlations between the frailties are then

$$\begin{aligned}
\mathrm{Corr}(Z_1, Z_2) &= \frac{k_0 + k_1}{\alpha} = 1 \\
&= \mathrm{Corr}(Z_3, Z_4) \\
\mathrm{Corr}(Z_1, Z_3) &= \frac{k_0}{\alpha} = \rho \\
&= \mathrm{Corr}(Z_1, Z_4) \\
&= \mathrm{Corr}(Z_2, Z_3) \\
&= \mathrm{Corr}(Z_2, Z_4).
\end{aligned}$$

So the correlation between Z_1 and Z_2 and between Z_3 and Z_4 is equal to one, as expected since member 1 and 2 and member 3 and 4 share all their frailties by construction.
Since $\rho = \frac{k_0}{\alpha}$, $k_0 = \rho\alpha = \frac{\rho}{\sigma^2}$ and since $\frac{k_0+k_1}{\alpha^2} = \sigma^2$, $k_1 = \frac{1}{\sigma^2} - k_0 = \frac{1-\rho}{\sigma^2}$. Using $k_0 \geqslant 0$ and $k_1 \geqslant 0$, the range of the correlation coefficient ρ is

$$0 \leqslant \rho \leqslant 1.$$

We further assume that given the frailties Z_1, Z_2, Z_3 and Z_4, the event times T_1, T_2, T_3 and T_4 are independent. Under this condition the marginal survival function $S(t_1, t_2, t_3, t_4)$ can be obtained from the conditional survival

function as follows:

$$\begin{aligned} S(t_1,t_2,t_3,t_4) &= \mathrm{E}(S(t_1,t_2,t_3,t_4|Z_1,Z_2,Z_3,Z_4)) \\ &= \mathrm{E}\left(S_1(t_1|Z_1)S_2(t_2|Z_2)S_3(t_3|Z_3)S_4(t_4|Z_4)\right) \\ &= \mathrm{E}\left(e^{-Z_1H_1(t_1)}e^{-Z_2H_2(t_2)}e^{-Z_3H_3(t_3)}e^{-Z_4H_4(t_4)}\right) \\ &= \mathrm{E}\Big(e^{-(Y_0+Y_1)H_1(t_1)}e^{-(Y_0+Y_1)H_2(t_2)} \\ &\quad e^{-(Y_0+Y_2)H_3(t_3)}e^{-(Y_0+Y_2)H_4(t_4)}\Big) \\ &= \mathrm{E}\Big(e^{-Y_0(H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4))} \\ &\quad e^{-Y_1(H_1(t_1)+H_2(t_2))}e^{-Y_2(H_3(t_3)+H_4(t_4))}\Big). \end{aligned} \tag{6.10}$$

Making use of the Laplace transform of the gamma distribution (1.10), the marginal survival function (6.10) can be written as

$$\begin{aligned} S(t_1,t_2,t_3,t_4) &= \left(1+\frac{H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4)}{\alpha}\right)^{-k_0} \\ &\quad \left(1+\frac{H_1(t_1)+H_2(t_2)}{\alpha}\right)^{-k_1}\left(1+\frac{H_3(t_3)+H_4(t_4)}{\alpha}\right)^{-k_1} \\ &= \left(1+\sigma^2(H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4))\right)^{\frac{-\rho}{\sigma^2}} \\ &\quad \left(1+\sigma^2(H_1(t_1)+H_2(t_2))\right)^{\frac{\rho-1}{\sigma^2}} \\ &\quad \left(1+\sigma^2(H_3(t_3)+H_4(t_4))\right)^{\frac{\rho-1}{\sigma^2}}. \end{aligned}$$

Using (6.3) the marginal survival function can be written as

$$S(t_1,t_2,t_3,t_4) = A^{-\rho/\sigma^2}B^{(\rho-1)/\sigma^2}C^{(\rho-1)/\sigma^2}, \tag{6.11}$$

with $A = S_1(t_1)^{-\sigma^2}+S_2(t_2)^{-\sigma^2}+S_3(t_3)^{-\sigma^2}+S_4(t_4)^{-\sigma^2}-3$, $B = S_1(t_1)^{-\sigma^2}+S_2(t_2)^{-\sigma^2}-1$ and $C = S_3(t_3)^{-\sigma^2}+S_4(t_4)^{-\sigma^2}-1$. For $\rho = 1$ the shared gamma frailty model is obtained, for $\rho = 0$ two bivariate correlated gamma frailty models are obtained.
The partial derivatives of the marginal survival function, needed to construct the likelihood, are

$$\begin{aligned} \frac{\partial S(t_1,t_2,t_3,t_4)}{\partial t_j} &= -S(t_1,t_2,t_3,t_4)S_j(t_j)^{-\sigma^2-1}f_j(t_j) \\ &\quad \left[\rho A^{-1}+(1-\rho)B^{-1}\right] \qquad \text{for } j=1,2 \end{aligned}$$

$$\frac{\partial S(t_1,t_2,t_3,t_4)}{\partial t_j} = -S(t_1,t_2,t_3,t_4)S_j(t_j)^{-\sigma^2-1}f_j(t_j) \Big[\rho A^{-1} + (1-\rho)C^{-1}\Big] \quad \text{for } j=3,4$$

$$\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} = S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1}f_j(t_j)f_k(t_k) \Big[\Big[\rho A^{-1} + (1-\rho)B^{-1}\Big]^2 + \rho\sigma^2 A^{-2} + (1-\rho)\sigma^2 B^{-2}\Big] \quad \text{for } j,k=1,2;\ j<k$$

$$\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} = S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1}f_j(t_j)f_k(t_k) \Big[\Big[\rho A^{-1} + (1-\rho)C^{-1}\Big]^2 + \rho\sigma^2 A^{-2} + (1-\rho)\sigma^2 C^{-2}\Big] \quad \text{for } j,k=3,4;\ j<k$$

$$\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} = S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1}f_j(t_j)f_k(t_k) \Big[\Big[\rho A^{-1} + (1-\rho)C^{-1}\Big]\Big[\rho A^{-1} + (1-\rho)B^{-1}\Big] + \rho\sigma^2 A^{-2}\Big] \quad \text{for } j=1,2; k=3,4$$

$$\frac{\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l} = -S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k)S_l(t_l))^{-\sigma^2-1} f_j(t_j)f_k(t_k)f_l(t_l) \Big\{\Big[\rho A^{-1} + (1-\rho)C^{-1}\Big]\Big[\Big[\rho A^{-1} + (1-\rho)B^{-1}\Big]^2 + \sigma^2\rho A^{-2} + (1-\rho)\sigma^2 B^{-2}\Big] + 2\rho\sigma^2 A^{-2}\Big[\rho A^{-1} + (1-\rho)B^{-1}\Big] + 2\sigma^4\rho A^{-3}\Big\} \quad \text{for } j=1; k=2; l=3,4$$

$$
\begin{aligned}
\frac{\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l} &= -S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k)S_l(t_l))^{-\sigma^2-1} \\
&\quad f_j(t_j)f_k(t_k)f_l(t_l) \\
&\quad \Big\{\big[\rho A^{-1}+(1-\rho)C^{-1}\big]\Big[\big[\rho A^{-1}+(1-\rho)C^{-1}\big] \\
&\quad \big[\rho A^{-1}+(1-\rho)B^{-1}\big]+\sigma^2\rho A^{-2}\Big] \\
&\quad +\big[\rho\sigma^2 A^{-2}+(1-\rho)\sigma^2 C^{-2}\big]\big[\rho A^{-1}+(1-\rho)B^{-1}\big] \\
&\quad +\big[\rho A^{-1}+(1-\rho)C^{-1}\big]\big[\rho\sigma^2 A^{-2}\big]+2\sigma^4\rho A^{-3}\Big\} \\
&\qquad\qquad \text{for } j=1,2; k=3; l=4
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^4 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l \partial t_m} &= S(t_1,t_2,t_3,t_4)\Big(S_j(t_j)S_k(t_k)S_l(t_l)S_m(t_m)\Big)^{-\sigma^2-1} \\
&\quad f_j(t_j)f_k(t_k)f_l(t_l)f_m(t_m) \\
&\quad \Big\{\big[\rho A^{-1}+(1-\rho)C^{-1}\big]\Big[\big[\rho A^{-1}+(1-\rho)C^{-1}\big] \\
&\quad \Big[\big[\rho A^{-1}+(1-\rho)B^{-1}\big]^2+\rho\sigma^2 A^{-2}+(1-\rho)\sigma^2 B^{-2}\Big] \\
&\quad +2\sigma^2\rho A^{-2}\big[\rho A^{-1}+(1-\rho)B^{-1}\big]+2\rho\sigma^4 A^{-3}\Big] \\
&\quad +\big[\rho\sigma^2 A^{-2}+(1-\rho)\sigma^2 C^{-2}\big]\Big[\big[\rho A^{-1}+(1-\rho)B^{-1}\big]^2 \\
&\quad +\rho\sigma^2 A^{-2}+(1-\rho)\sigma^2 B^{-2}\Big] \\
&\quad +\big[\rho A^{-1}+(1-\rho)C^{-1}\big]\Big[2\big[\rho A^{-1}+(1-\rho)B^{-1}\big]\rho\sigma^2 A^{-2} \\
&\quad +2\rho\sigma^4 A^{-3}\Big]+4\rho\sigma^4 A^{-3}\big[\rho A^{-1}+(1-\rho)B^{-1}\big] \\
&\quad +2\rho\sigma^2 A^{-2}\big[\rho\sigma^2 A^{-2}\big]+6\rho\sigma^6 A^{-4}\Big\}. \\
&\qquad\qquad \text{for } j=1; k=2; l=3; m=4
\end{aligned}
$$

The loglikelihood is given by (6.9). If a parametric distribution is chosen for the baseline hazard, this closed form expression of the loglikelihood can be maximized with respect to the unknown parameters. Standard errors can be obtained from the inverse of the observed information matrix. Representation (6.11) of the joint survival function also allows a semiparametric two-stage copula approach.

6.3.3 The fourdimensional correlated gamma frailty model with correlation ρ_1 and ρ_2 between different frailties (model 4)

Next, the model is extended in the sense that the correlation between the frailties of the first two members of the cluster and between the last two members of the cluster is not necessarily equal to one, but can take a value ρ_1. The correlations between other frailty-pairs is then equal to ρ_2. Figure 6.3 visualizes the construction of the frailties and the correlation structure between them for the model described in this section. The individual frail-

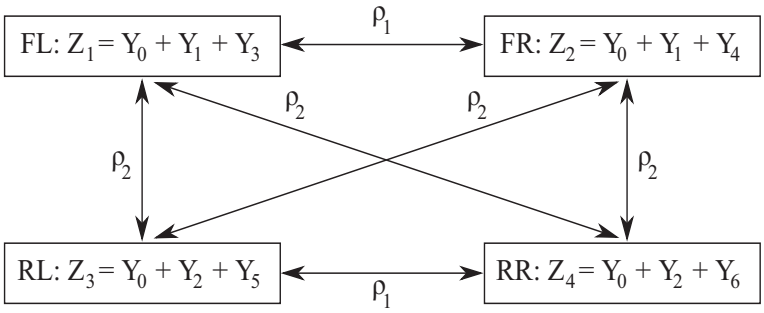

Figure 6.3: Construction of the frailties and the pairwise correlation structure for the fourdimensional correlated frailty model with correlations between the frailties equal to ρ_1 or ρ_2. Z_1, Z_2, Z_3 and Z_4 represent the frailty for the front left (FL), front right (FR), rear left (RL) and rear right (RR) udder quarter, respectively.

ties will be constructed using six independent gamma-distributed random variables Y_0, Y_1, Y_2, Y_3, Y_4, Y_5 and Y_6 with parameters (k_l, α_l), l=0,. . . ,6, respectively. Y_0 will represent the common part of the frailty for each member of the cluster. Y_1 and Y_2 are common for only two members of the cluster, for example Y_1 is common for the first two members and Y_2 is common for the other two members. Y_3, Y_4, Y_5 and Y_6 represent truly individual heterogeneity. Since there is a frailty-part that represents individual heterogeneity the model resembles more the originally proposed correlated frailty model (Yashin et al., 1995), but it is extended by a frailty-part that is only common to some of the members in the cluster. We again assume that the scale parameters of the gamma-distributed random variables Y_0, Y_1, Y_2, Y_3, Y_4, Y_5 and Y_6 are the same, i.e., $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha$. We further assume that the shape parameters for the distribution of Y_1 and Y_2

are the same, i.e., $k_1 = k_2$ and that the shape parameters for the distribution of Y_3, Y_4, Y_5 and Y_6 are the same, i.e., $k_3 = k_4 = k_5 = k_6$. We will use k_2 for the shape parameter of Y_3, Y_4, Y_5 and Y_6. The frailties are then given by

$$\begin{aligned} Z_1 &= Y_0 + Y_1 + Y_3 \sim \text{gamma}(k_0 + k_1 + k_2, \alpha) \\ Z_2 &= Y_0 + Y_1 + Y_4 \sim \text{gamma}(k_0 + k_1 + k_2, \alpha) \\ Z_3 &= Y_0 + Y_2 + Y_5 \sim \text{gamma}(k_0 + k_1 + k_2, \alpha) \\ Z_4 &= Y_0 + Y_2 + Y_6 \sim \text{gamma}(k_0 + k_1 + k_2, \alpha). \end{aligned}$$

We further add the restriction $k_0 + k_1 + k_2 = \alpha$. The frailties therefore have mean 1:

$$\text{E}(Z_1) = \text{E}(Z_2) = \text{E}(Z_3) = \text{E}(Z_4) = \frac{k_0 + k_1 + k_2}{\alpha} = 1$$

and we denote their variance by σ^2

$$\text{V}(Z_1) = \text{V}(Z_2) = \text{V}(Z_3) = \text{V}(Z_4) = \frac{k_0 + k_1 + k_2}{\alpha^2} = \frac{1}{\alpha} = \sigma^2.$$

The covariances between the frailties can be obtained as follows:

$$\begin{aligned} \text{Cov}(Z_1, Z_2) &= \text{Cov}(Y_0 + Y_1 + Y_3, Y_0 + Y_1 + Y_4) = V(Y_0 + Y_1) = \frac{k_0 + k_1}{\alpha^2} \\ \text{Cov}(Z_3, Z_4) &= \text{Cov}(Y_0 + Y_2 + Y_5, Y_0 + Y_2 + Y_6) = V(Y_0 + Y_2) = \frac{k_0 + k_1}{\alpha^2} \\ \text{Cov}(Z_i, Z_j) &= \text{Cov}(Y_0 + Y_1 + Y_{i+2}, Y_0 + Y_2 + Y_{j+2}) = V(Y_0) = \frac{k_0}{\alpha^2}, \end{aligned}$$

for i=1,2; j=3,4. The correlations between the frailties are then

$$\begin{aligned} \text{Corr}(Z_1, Z_2) &= \frac{k_0 + k_1}{\alpha} = \rho_1 \\ &= \text{Corr}(Z_3, Z_4) \\ \text{Corr}(Z_i, Z_j) &= \frac{k_0}{\alpha} = \rho_2, \end{aligned}$$

for i=1,2; j=3,4. The shape parameters k_0, k_1 and k_2 can be expressed as:

$$\begin{aligned} \rho_2 &= \frac{k_0}{\alpha} \quad \Rightarrow \quad k_0 = \rho_2 \alpha = \frac{\rho_2}{\sigma^2} \\ \rho_1 &= \frac{k_0 + k_1}{\alpha} \quad \Rightarrow \quad k_1 = \rho_1 \alpha - k_0 = \frac{\rho_1 - \rho_2}{\sigma^2} \\ 1 &= \frac{k_0 + k_1 + k_2}{\alpha} \\ &\Rightarrow k_2 = \alpha - k_0 - k_1 = \frac{1}{\sigma^2} - \frac{\rho_2}{\sigma^2} - \frac{\rho_1 - \rho_2}{\sigma^2} = \frac{1 - \rho_1}{\sigma^2}. \end{aligned}$$

Since the shape parameters need to be positive, the following restrictions on the correlations hold:

$$\begin{aligned} k_0 \geqslant 0 &\Rightarrow \rho_2 \geqslant 0 \\ k_1 \geqslant 0 &\Rightarrow \rho_1 \geqslant \rho_2 \\ k_2 \geqslant 0 &\Rightarrow \rho_1 \leqslant 1 \end{aligned}$$

or summarizing

$$0 \leqslant \rho_2 \leqslant \rho_1 \leqslant 1.$$

We again assume that given the frailties Z_1, Z_2, Z_3 and Z_4, the event times T_1, T_2, T_3 and T_4 are independent. Under this condition the marginal survival function $S(t_1, t_2, t_3, t_4)$ can be obtained from the conditional survival function as follows:

$$\begin{aligned} S(t_1,t_2,t_3,t_4) &= \mathrm{E}\left(S(t_1,t_2,t_3,t_4|Z_1,Z_2,Z_3,Z_4)\right) \\ &= \mathrm{E}\left(S_1(t_1|Z_1)S_2(t_2|Z_2)S_3(t_3|Z_3)S_4(t_4|Z_4)\right) \\ &= \mathrm{E}\left(e^{-Z_1H_1(t_1)}e^{-Z_2H_2(t_2)}e^{-Z_3H_3(t_3)}e^{-Z_4H_4(t_4)}\right) \\ &= \mathrm{E}\Big(e^{-(Y_0+Y_1+Y_3)H_1(t_1)}e^{-(Y_0+Y_1+Y_4)H_2(t_2)} \\ &\quad e^{-(Y_0+Y_2+Y_5)H_3(t_3)}e^{-(Y_0+Y_2+Y_6)H_4(t_4)}\Big) \\ &= \mathrm{E}\Big(e^{-Y_0(H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4))} \\ &\quad e^{-Y_1(H_1(t_1)+H_2(t_2))}e^{-Y_2(H_3(t_3)+H_4(t_4))} \\ &\quad e^{-Y_3H_1(t_1)}e^{-Y_4H_2(t_2)}e^{-Y_5H_3(t_3)}e^{-Y_6H_4(t_4)}\Big) \end{aligned} \tag{6.12}$$

Making use of the Laplace transform of the gamma distribution (1.10), the marginal survival function (6.12) can be written as

$$\begin{aligned} S(t_1,t_2,t_3,t_4) &= \left(1+\frac{H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4)}{\alpha}\right)^{-k_0} \\ &\quad \left(1+\frac{H_1(t_1)+H_2(t_2)}{\alpha}\right)^{-k_1}\left(1+\frac{H_3(t_3)+H_4(t_4)}{\alpha}\right)^{-k_1} \\ &\quad \left(1+\frac{H_1(t_1)}{\alpha}\right)^{-k_2}\left(1+\frac{H_2(t_2)}{\alpha}\right)^{-k_2} \\ &\quad \left(1+\frac{H_3(t_3)}{\alpha}\right)^{-k_2}\left(1+\frac{H_4(t_4)}{\alpha}\right)^{-k_2} \end{aligned}$$

$$
\begin{aligned}
= \quad & \left(1+\sigma^2(H_1(t_1)+H_2(t_2)+H_3(t_3)+H_4(t_4))\right)^{-\rho_2/\sigma^2} \\
& \left(1+\sigma^2(H_1(t_1)+H_2(t_2))\right)^{(\rho_2-\rho_1)/\sigma^2} \\
& \left(1+\sigma^2(H_3(t_3)+H_4(t_4))\right)^{(\rho_2-\rho_1)/\sigma^2} \\
& \left(1+\frac{H_1(t_1)}{\alpha}\right)^{(\rho_1-1)/\sigma^2} \left(1+\frac{H_2(t_2)}{\alpha}\right)^{(\rho_1-1)/\sigma^2} \\
& \left(1+\frac{H_3(t_3)}{\alpha}\right)^{(\rho_1-1)/\sigma^2} \left(1+\frac{H_4(t_4)}{\alpha}\right)^{(\rho_1-1)/\sigma^2}
\end{aligned}
$$

Using (6.3) the marginal survival function can be written as

$$
\begin{aligned}
S(t_1,t_2,t_3,t_4) \quad = \quad & A^{-\rho_2/\sigma^2} B^{(\rho_2-\rho_1)/\sigma^2} C^{(\rho_2-\rho_1)/\sigma^2} \\
& (S_1(t_1)S_2(t_2)S_3(t_3)S_4(t_4))^{1-\rho_1}\,, \qquad (6.13)
\end{aligned}
$$

with $A = S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} + S_3(t_3)^{-\sigma^2} + S_4(t_4)^{-\sigma^2} - 3$, $B = S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1$ and $C = S_3(t_3)^{-\sigma^2} + S_4(t_4)^{-\sigma^2} - 1$. For $\rho_1 = 1$ the fourdimensional correlated gamma frailty model with shared and correlated frailties (6.11) is obtained. For $\rho_1 = \rho_2 = \rho$ the fourdimensional correlated gamma frailty model with equal correlation between the frailties (6.7) is obtained. Taking $\rho_1 = \rho_2 = 1$ leads to the shared gamma frailty model.
The partial derivatives of the marginal survival function, needed to construct the likelihood, are

$$
\begin{aligned}
\frac{\partial S(t_1,t_2,t_3,t_4)}{\partial t_j} \quad = \quad & -S(t_1,t_2,t_3,t_4) S_j(t_j)^{-\sigma^2-1} f_j(t_j) \\
& \left[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\right] \\
& \text{for } j = 1,2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial S(t_1,t_2,t_3,t_4)}{\partial t_j} \quad = \quad & -S(t_1,t_2,t_3,t_4) S_j(t_j)^{-\sigma^2-1} f_j(t_j) \\
& \left[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\right] \\
& \text{for } j = 3,4
\end{aligned}
$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} &= S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1} f_j(t_j) f_k(t_k) \\ &\Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\ &\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\ &+\rho_2\sigma^2 A^{-2} + \sigma^2(\rho_1-\rho_2)B^{-2}\Big] \\ &\qquad\qquad \text{for } j,k = 1,2;\ j<k\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} &= S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1} f_j(t_j) f_k(t_k) \\ &\Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\ &\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\ &+\rho_2\sigma^2 A^{-2} + \sigma^2(\rho_1-\rho_2)C^{-2}\Big] \\ &\qquad\qquad \text{for } j,k = 3,4;\ j<k\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k} &= S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k))^{-\sigma^2-1} f_j(t_j) f_k(t_k) \\ &\Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\ &\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\ &+\rho_2\sigma^2 A^{-2}\Big] \qquad\qquad \text{for } j = 1,2; k = 3,4\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l} &= -S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k)S_l(t_l))^{-\sigma^2-1} \\
&\quad f_j(t_j)f_k(t_k)f_l(t_l) \\
&\quad \Big\{\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_l(t_l)^{\sigma^2}\big] \\
&\quad \Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\sigma^2\rho_2 A^{-2} + (\rho_1-\rho_2)\sigma^2 B^{-2}\Big] \\
&\quad +\rho_2\sigma^2 A^{-2}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\rho_2\sigma^2 A^{-2}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad +2\sigma^4\rho_2 A^{-3}\Big\} \qquad \text{for } j=1; k=2; l=3,4
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^3 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l} &= -S(t_1,t_2,t_3,t_4)(S_j(t_j)S_k(t_k)S_l(t_l))^{-\sigma^2-1} \\
&\quad f_j(t_j)f_k(t_k)f_l(t_l) \\
&\quad \Big\{\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_l(t_l)^{\sigma^2}\big] \\
&\quad \Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\sigma^2\rho_2 A^{-2}\Big] \\
&\quad +\big[\rho_2\sigma^2 A^{-2} + \sigma^2(\rho_1-\rho_2)C^{-2}\big] \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\rho_2\sigma^2 A^{-2}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad +2\sigma^4\rho_2 A^{-3}\Big\} \qquad \text{for } j=1,2; k=3; l=4
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^4 S(t_1,t_2,t_3,t_4)}{\partial t_j \partial t_k \partial t_l \partial t_m} &= S(t_1,t_2,t_3,t_4)\Big(S_j(t_j)S_k(t_k)S_l(t_l)S_m(t_m)\Big)^{-\sigma^2-1} \\
&\quad f_j(t_j)f_k(t_k)f_l(t_l)f_m(t_m) \\
&\quad \Bigg\{\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_m(t_m)^{\sigma^2}\big] \\
&\quad \Bigg[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_l(t_l)^{\sigma^2}\big] \\
&\quad \Bigg[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\rho_2\sigma^2 A^{-2} + (\rho_1-\rho_2)\sigma^2 B^{-2}\Big] \\
&\quad +\sigma^2\rho_2 A^{-2}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\sigma^2\rho_2 A^{-2}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad +2\rho_2\sigma^4 A^{-3}\Big] + \big[\rho_2\sigma^2 A^{-2} + (\rho_1-\rho_2)\sigma^2 C^{-2}\big] \\
&\quad \Bigg[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big] \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +\rho_2\sigma^2 A^{-2} + (\rho_1-\rho_2)\sigma^2 B^{-2}\Big] \\
&\quad +\big[\rho_2 A^{-1} + (\rho_1-\rho_2)C^{-1} + (1-\rho_1)S_l(t_l)^{\sigma^2}\big] \\
&\quad \Big[\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big]\rho_2\sigma^2 A^{-2} \\
&\quad +\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_k(t_k)^{\sigma^2}\big]\rho_2\sigma^2 A^{-2} \\
&\quad +2\rho_2\sigma^4 A^{-3}\Big] + 2\rho_2\sigma^4 A^{-3} \\
&\quad \big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + (1-\rho_1)S_j(t_j)^{\sigma^2}\big] \\
&\quad +2\big[\rho_2\sigma^2 A^{-2}\big]^2 \\
&\quad +2\rho_2\sigma^4 A^{-3}\big[\rho_2 A^{-1} + (\rho_1-\rho_2)B^{-1} + \\
&\quad (1-\rho_1)S_k(t_k)^{\sigma^2}\big] + 6\rho_2\sigma^6 A^{-4}\Bigg\} \\
&\qquad\qquad \text{for } j=1; k=2; l=3; m=4
\end{aligned}
$$

The loglikelihood is given by (6.9). If a parametric distribution is chosen for the baseline hazard, this closed form expression of the loglikelihood can be maximized with respect to the unknown parameters. Standard errors can

be obtained from the inverse of the observed information matrix. Representation (6.13) of the joint survival function also allows a semiparametric two-stage copula approach.

6.4 Analysis of the mastitis data

The proposed models are applied to the time to infection with *C. bovis* data set. All programs are written in R. A Weibull distribution is assumed for the baseline hazard. We investigate the effect of the udder quarter location (front or rear), an udder quarter level covariate, and the effect of parity (multiparous versus primiparous), a between cow covariate, on the time to infection with *C. bovis*. Special interest is now in the estimates of the correlation parameters. Since the frailty is defined as an individual susceptibility to experience the event, i.e., to get infected with *C. bovis*, the correlations between the different frailties describe the correlations between the susceptibility of the different udder quarters to get infected. We also compare the estimates obtained in the proposed models with estimates obtained from fitting a shared gamma frailty model (model 1) to the time to infection with *C. bovis* data set. In the shared gamma frailty model the correlation between the frailties of the different udder quarters is equal to one.
Different models were fitted to the data. Since the models are not nested within each other, comparison of the models is based on the Akaike Information Criterion (AIC). To calculate the AIC we use the formula AIC $= -2 \log L + 2 \times$ (number of parameters) (Izumi and Ohtaki, 2004).
In the model with the correlation equal to one or ρ (model 3) and in the model with the correlation equal to ρ_1 or ρ_2 (model 4) there is a restriction on the magnitude of the correlation parameters: $\rho \leqslant 1$ in model 3 and $\rho_2 \leqslant \rho_1$ in model 4. First we investigate whether it is more sensible to assume that the correlation between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters is bigger than the correlation between the frailties of the two left udder quarters and between the frailties of the two right udder quarters or whether it is more sensible to make the opposite assumption. The AIC for the models with the strongest correlation between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters is smaller (7806.464 for model 3 and 7807.008 for model 4 versus 7843.602 for model 3 and 7845.18 for model 4 when the opposite assumption is made). Therefore, we will proceed with the assumption that the correlation is strongest between

the frailties of the two front udder quarters and between the frailties of the two rear udder quarters.
The parameter estimates and their standard errors for the different models are given in Table 6.1. The parameter estimates in all models are compa-

Table 6.1: Parameter estimates (Est) and their standard errors (SE) for the shared gamma frailty model (model 1) and three correlated gamma frailty models (model 2, 3 and 4) with parity ($\hat{\beta}_p$ is the effect of a multiparous cow) and udder quarter location ($\hat{\beta}_l$ is the effect of the rear udder quarter) as covariates and Weibull baseline hazard.

	Shared (model 1) Est (SE)	Corr=ρ (model 2) Est (SE)	Corr=1 or ρ (model 3) Est (SE)	Corr=ρ_1 or ρ_2 (model 4) Est (SE)
θ	3.846 (0.223)	3.914 (0.272)	4.132 (0.250)	4.338 (0.332)
λ	0.138 (0.016)	0.138 (0.017)	0.143 (0.018)	0.145 (0.019)
γ	1.981 (0.040)	1.991 (0.047)	2.038 (0.045)	2.074 (0.059)
ρ	-	0.993 (0.015)	0.954 (0.017)	-
ρ_1	-	-	-	0.984 (0.016)
ρ_2	-	-	-	0.936 (0.024)
β_l	-0.275 (0.050)	-0.277 (0.050)	-0.288 (0.053)	-0.294 (0.054)
β_p	0.866 (0.143)	0.875 (0.146)	0.881 (0.149)	0.907 (0.156)
AIC	7783.318	7785.114	7775.846	7776.732

rable. The estimate for γ in each model is above 1, therefore, the hazard is increasing with time. The conclusions concerning the covariates are also the same in all models. Since model 3 has the lowest value for the AIC, model 3 is selected as the best fitting model. We discuss the effect of the covariates based on the estimates obtained in this model. The rear udder quarters have a significantly lower hazard of infection than the front udder quarters, with hazard ratio (HR) = 0.75 (95% confidence interval (CI) [0.68;0.83]). The hazard of infection for multiparous cows was significantly higher compared to heifers, HR = 2.41, 95% CI [1.80;3.23].
The estimate for θ is lowest in the shared frailty model. This makes sense, since in the shared frailty model θ represents the variance of frailties that represent heterogeneity due to only common unobserved risk factors, while in the other (correlated) frailty models the frailties represent heterogeneity due to a combination of common and individual unobserved risk factors.

Comparing model 1 and model 2, it can be seen that the estimate for the parameter ρ is lower in model 2 (ρ is equal to one in the shared frailty model), while the estimate for θ is a bit higher in model 2 than in model 1. This relation between ρ and θ is already described for the bivariate case in Wienke et al. (2005a).
The correlations between the frailties of the different udder quarters are high in all models. In model 2 the correlation between the frailties of the different udder quarters is the same and equal to 0.993. The AIC for model 2 is higher than the AIC for the shared frailty model implying that the extra parameter ρ does not improve the fit and the shared frailty model with ρ equal to one is adequate. However introducing the possibility of a different correlation between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters on the one hand and between the frailties of the two left udder quarters and between the frailties of the two right udder quarters on the other hand, significantly improves the fit, resulting in a lower value for the AIC, despite the fact that an extra parameter needs to be introduced in the model. Comparing model 3 and model 4 we can conclude that it is good to suppose a different correlation between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters on the one hand and between the frailties of the two left udder quarters and between the frailties of the two right udder quarters on the other hand, but that the strongest correlation can be assumed to be equal to one.

6.5 The fourdimensional correlated gamma frailty model for interval-censored data

In the previous section different correlated gamma frailty models were fitted to fourdimensional interval-censored data by imputing the midpoint of the interval as an exact event time. In this section the interval-censored nature of the data is taken into account. The fourdimensional distribution of frailty and the correlation structure is obtained in the same way as in the previous section for all models. The joint survival function is also obtained as the expectation of the conditional survival function and takes the form (6.7), (6.11) or (6.13) for model 2, 3, 4, respectively.
In Section 4.4 the loglikelihood for the copula model for fourdimensional interval-censored data is constructed. Since the likelihood is expressed in terms of the joint survival function, it can also be applied here with the joint survival function given by (6.7), (6.11) or (6.13) for model 2, 3, 4,

respectively. The loglikelihood (6.9) therefore needs to be replaced by

$$\begin{aligned}\log L(\boldsymbol{\zeta}) \quad = \quad & \sum_{i=1}^{k}\Bigg[\Delta_i \log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4}) \\ & + \sum_{j=1}^{4}\left[\Delta_i(j)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij})\right] \\ & + \sum_{j\neq k}\left[\Delta_i(j,k)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij},u_{ik})\right] \\ & + \sum_{j\neq k;j\neq l;k\neq l}\left[\Delta_i(j,k,l)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{ij},u_{ik},u_{il})\right] \\ & + \Delta_i(1,2,3,4)\log L_{i,\boldsymbol{\delta_i}}(l_{i1},l_{i2},l_{i3},l_{i4},u_{i1},u_{i2},u_{i3},u_{i4})\Bigg],\end{aligned}$$

with $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, $\boldsymbol{\xi}$ containing the parameters of the baseline hazard, $\boldsymbol{\delta_i} = (\delta_{i1}, \delta_{i2}, \delta_{i3}, \delta_{i4})$, with δ_{ij}, $j = 1, 2, 3, 4$, equal to 1 if the j^{th} member of the i^{th} cluster is interval-censored and equal to 0 if the j^{th} member of the i^{th} cluster is right-censored, Δ_i, $\Delta_i(j)$, $\Delta_i(j,k)$, $\Delta_i(j,k,l)$ and $\Delta_i(1,2,3,4)$ defined as in (4.6) and k the number of clusters. For the different likelihood contributions $L_{i,\boldsymbol{\delta_i}}$ we refer to Section 4.4, with the joint survival function $S(p,q,r,s)$ given by (6.7), (6.11) and (6.13) for model 2, 3 and 4, respectively.
Parameter estimates and their standard errors for the different models are given in Table 6.2. Conclusions concerning the effect of covariates, the correlation structure in the data and the best fitting model are the same as when imputation of the midpoint is used.

6.6 Conclusions

In this chapter we proposed three fourdimensional correlated gamma frailty models to model the correlation structure between the frailties of the four udder quarters of a dairy cow. Model 4 could even be extended to a model with three different correlations: a correlation equal to ρ_1 between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters, a correlation equal to ρ_2 between the frailties of the two left udder quarters and between the frailties of the two right udder quarters and a correlation equal to ρ_3 between the frailties of the udder quarters that are positioned diagonally from each other. However, the model would become quite complex and based on the results obtained in model 4, we believe

Table 6.2: Parameter estimates (Est) and their standard errors (SE) for a shared gamma frailty model (model 1) and three correlated gamma frailty models (model 2, 3 and 4) with parity ($\hat{\beta}_p$ is the effect of a multiparous cow) and udder quarter location ($\hat{\beta}_l$ is the effect of the rear udder quarter) as covariates and Weibull baseline hazard.

	Shared (model 1) Est (SE)	Corr=ρ (model 2) Est (SE)	Corr=1 or ρ (model 3) Est (SE)	Corr=ρ_1 or ρ_2 (model 4) Est (SE)
θ	3.842 (0.224)	3.883 (0.271)	4.119 (0.250)	4.280 (0.333)
λ	0.137 (0.017)	0.138 (0.017)	0.142 (0.018)	0.144 (0.019)
γ	1.984 (0.042)	1.991 (0.049)	2.041 (0.047)	2.070 (0.062)
ρ	-	0.996 (0.015)	0.956 (0.016)	-
ρ_1	-	-	-	0.987 (0.016)
ρ_2	-	-	-	0.942 (0.025)
β_l	-0.276 (0.050)	-0.278 (0.051)	-0.290 (0.054)	-0.295 (0.055)
β_p	0.867 (0.143)	0.872 (0.146)	0.882 (0.150)	0.903 (0.156)
AIC	11320.83	11322.76	11312.22	11313.53

introducing an extra parameter would not improve the fit for the infection with *C. bovis* data. Based on the AIC, a model that allows a different correlation between the frailties of the two front udder quarters and between the frailties of the two rear udder quarters on the one hand and between the frailties of the two left udder quarters and between the frailties of the two right udder quarters on the other hand, with the extra simplification that the strongest correlation is equal to one, provides the best fit to the infection with *C. bovis* data.

The proposed models have several advantages. By assuming a gamma distribution for the frailties, the frailties can be integrated out from the conditional likelihood and a closed form expression for the marginal likelihood is obtained which can then be maximized by traditional estimation methods (maximum likelihood estimation) to obtain parameter estimates and their standard errors. The representation of the joint survival function in terms of the marginal survival functions and the correlation parameters allows a semiparametric two-stage copula estimation approach and makes the model flexible. It is however important to keep in mind that the standard errors of the estimates of the correlation parameters obtained in the second stage

do not include the error that is created by using a nonparametric or semiparametric estimator for the marginal survival functions in the first stage. Therefore, other methods such as, for example, a bootstrap approach should be used to obtain standard errors of the parameter estimates. Using a fully parametric approach has the advantage that standard errors can be obtained directly from the information matrix.
The proposed models also have disadvantages, which may present limitations in practical situations. The construction of the fourdimensional frailty distribution imposes constraints on the correlations between the frailty variables. This restricts the possible range of correlation structures between the frailties Z_1, Z_2, Z_3 and Z_4. However, if we restrict to the models presented in this chapter, restrictions are reasonable for practical situations. It is further important to realize that the parameter ρ in model 2 and model 3 and the parameters ρ_1 and ρ_2 in model 4 describe the correlation between the frailties and not the correlation between the event times. Lindeboom and Van Den Berg (1994) investigate the relationship between the correlation between frailties and the correlation between event times. They derive explicit expressions for the correlation between the event times in the special case of a constant baseline hazard function. Unfortunately, such explicit results are not available for more general situations.

Chapter 7

Conclusions and further research

Mastitis, the inflammation of the udder of a dairy cow, is economically the most important disease in the dairy sector of the western world because udder infections are closely associated with reduced milk yield and milk quality (Seegers et al., 2003). Therefore, mastitis control is an important component of dairy herd health programs. As a consequence, it is important to interpret the results from experimental or observational studies, carried out to investigate, for example, possible risk factors for mastitis or the infectiousness of mastitis in a correct way. To accomplish this goal, not only an adequate experimental design is important; a proper statistical analysis of the observed data is also essential. The statistical model should exploit the information in the data to its full extent and should model the specific data structure correctly. For the mastitis data it is therefore important to use statistical models that take into account the clustering in the data and the interval-censored nature of the data simultaneously.

If the only goal of a mastitis study is to investigate the effect of covariates, the marginal model provides consistent estimates. An estimate of a parameter is consistent if the difference between the estimate and the actual parameter goes to an infinitesimally small value ϵ ($\epsilon > 0$) with probability 1 for the sample size n going to infinity. The likelihood-based estimates of the variance of the covariate effects, usually provided in commercial software packages, are however not consistent. To investigate and test the effect of the covariates correctly, for example by constructing confidence intervals, it is important that correct variances are used. Too small variances lead to too narrow confidence intervals and therefore type I errors larger than the proposed α. Too large variances lead to too wide confidence intervals, maybe missing a significant covariate effect. Correct variances can be obtained by using the grouped jackknife technique. The conclusions concerning the effect of the location covariate (β_l) and the parity covariate (β_p) in the mastitis data are not altered for any of the considered bacteria if the naïve (likelihood-based) estimate of the variance would be used. However, wrong conclusions could be drawn in other data sets, leading to false guidelines for the practitioner.

If interest is only in the covariate effects, the fixed effects model is another option to model clustered, interval-censored data. However, we recommend using the marginal model over the fixed effects model because of different disadvantages of the fixed effects model. First, the parameter estimate for λ (the scale parameter of the Weibull distribution) corresponds to the parameter for one particular cow only. Therefore, the censoring status of the

udder quarters in that cow determines whether this parameter can be estimated or not (if all udder quarters are censored, the parameter can not be estimated). However, most software packages still provide an arbitrary low value for the estimate and standard error of the parameter even if it can not be estimated. Second, the software also provides parameter estimates and their variances for all the fixed cow effects which are not really of interest. If the number of fixed cow effects is large, the software is sometimes not able to fit the model due to insufficient memory. Third, it is impossible to obtain estimates for cow level covariates in the fixed effects model, because there is complete confounding between the cow fixed effects and the cow level covariates. Nevertheless, statistical software packages typically provide a senseless estimate for these effects.

In this thesis different models are proposed that can be used if interest is not only in the effect of covariates, but also in the correlation in the data and/or in the possible correlation structures in the data. The proposed models also take into account the interval-censored nature of the mastitis data.
The first model discussed is the Clayton copula model. The copula model can only be used for data consisting of small clusters of equal size. If a parametric distribution is chosen for the marginal survival functions, either a two-stage estimation approach or a one-stage estimation approach is possible. If the marginal survival functions are estimated nonparametrically or semiparametrically, a two-stage estimation approach has to be used. The standard error for the estimate of θ (the correlation parameter in the copula model), obtained in the second stage of a two-stage estimation approach, does not take into consideration the uncertainty related to the estimation of the marginal survival functions in the first stage. Therefore, this standard error is incorrect, especially with small sample sizes, and other methods such as, for example, a bootstrap approach should be used to obtain a standard error of the estimate for θ. Sun et al. (2006) proof for bivariate data that the estimate for θ is consistent and asymptotically normal under certain regularity conditions if the marginal survival functions are estimated nonparametrically in the first stage of a two-stage estimation procedure. Investigating the properties of $\hat{\theta}$ in four dimensions is a topic of further research. For the infection with *Staph. aureus*, *Strep. dysgalactiae* and *Strep. uberis* data sets parameter estimates and their standard errors are very similar in the one-stage and parametric two-stage approach. Only the standard error for the estimate of γ (the shape parameter of the Weibull distribution) in the one-stage approach is slightly higher than the jackknife estimate obtained in the two-stage approach. For the infection with *C. bovis* data set the estimates

of θ, β_l and β_p differ depending on the approach used. The variance for the estimate of β_p (the effect of a multiparous cow) in the one-stage approach is lower than the jackknife estimate obtained in the two-stage approach. The one-stage approach is recommended in a fully parametric copula model because it leads to correct variance estimators without a requirement for bootstrap or jackknife techniques.
To interpret the parameter θ in the fourdimensional Clayton copula model its relationship with Kendall's τ is used. Kendall's τ is a global measure of correlation, defined in the bivariate case as

$$\tau = \mathrm{P}\left((T_{i1} - T_{k1})(T_{i2} - T_{k2}) > 0\right) - \mathrm{P}\left((T_{i1} - T_{k1})(T_{i2} - T_{k2}) < 0\right),$$

with (T_{i1}, T_{i2}), (T_{k1}, T_{k2}) the event times in two randomly chosen pairs. Values for τ are between -1 and 1, 1 corresponding to a perfect correlation, -1 meaning a perfect inverse correlation. If Kendall's τ is equal to 0, the event times are independent. A multivariate definition of Kendall's τ can be found in Nelsen (1996). The relationship between Kendall's τ and θ in the Clayton copula is given by $\tau = \theta/(\theta + 2)$. Using this relationship θ can be interpreted as a measure of the correlation between the infection times within a cow. The correlation in the copula model is necessarily positive. If the correlation is high, an infection of one udder quarter can easily evolve in infection of the other udder quarters. Therefore, it is important to take preventive measures to keep the other udder quarters infection free if one of the udder quarters of a cow is infected.

We further proposed a shared gamma frailty model for interval-censored data in a parametric setting. This model allows different and large cluster sizes. Assuming a gamma distribution for the frailties a closed form expression for the marginal likelihood can be obtained which can be maximized to obtain parameter estimates. Exact expressions for the second derivatives of the likelihood and thus estimates for the variances of the parameter estimates can be obtained by inverting the matrix of second derivatives or can be obtained from the Hessian matrix at the end of the maximization procedure. Although the proposed methodology and the R-program is valid for any cluster size, its practical applicability at the moment is restricted to data with at most eleven events in a cluster due to insufficient memory. The Kronecker product in the program needs to be reprogrammed in a less memory consuming way so that data with more than eleven events in a cluster can be fitted.
A direct interpretation of the parameter θ in the shared gamma frailty model

is not easy. θ is the variance between the frailties, thus representing between-cow variability. Variance between cows induces correlation within the cow. Thus, the larger the variance between the cows, the larger the correlation within the cow. Figure 7.1 depicts the relationship between Kendall's τ and θ. It can be seen that for values of θ between zero and two the increase in Kendall's τ is steep, but that for larger values of θ the increase in Kendall's τ levels off. Therefore, correlation is high in all data sets considered in this thesis with τ equal to 0.74, 0.66, 0.62 and 0.74 for infections with *Staph. aureus*, *C. bovis*, *Strep. dysgalactiae* and *Strep. uberis*, respectively, but a seemingly large difference in the value of θ (θ equal to 3 or 5) does not translate in a large difference for the value of Kendall's τ.
Another reason why the interpretation of the parameter θ is not straightforward in the shared gamma frailty model is the fact that the frailty operates on the hazard of infection in a multiplicative way, while interest is usually in the effect of the frailty on the infection times. Therefore, it is meaningful to

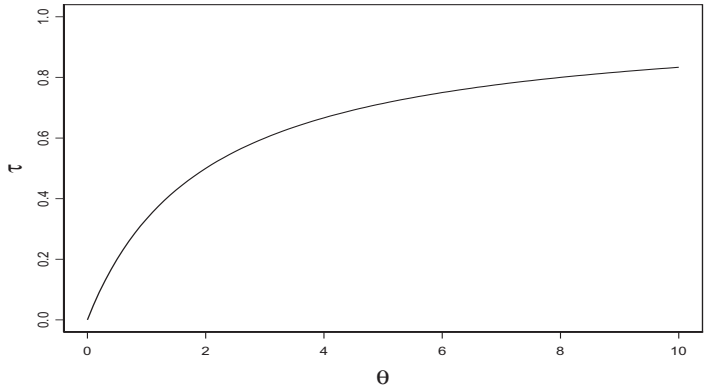

Figure 7.1: Relationship between τ and θ with $\tau = \theta/(\theta + 2)$

investigate how the frailties influence the median time to infection, a quantity that has a biological meaning (Duchateau and Janssen, 2005). Density functions for the median time to infection with *C. bovis* are depicted in Figure 7.2 for different values of θ. The larger the parameter θ the flatter the density function, the smaller the parameter θ the higher the peak. Figure 7.2 shows that the biggest differences in the density function for the median time to infection are obtained for values of θ between 0 and 1. For values of θ larger than 1, differences are less pronounced. Therefore, the effect of the

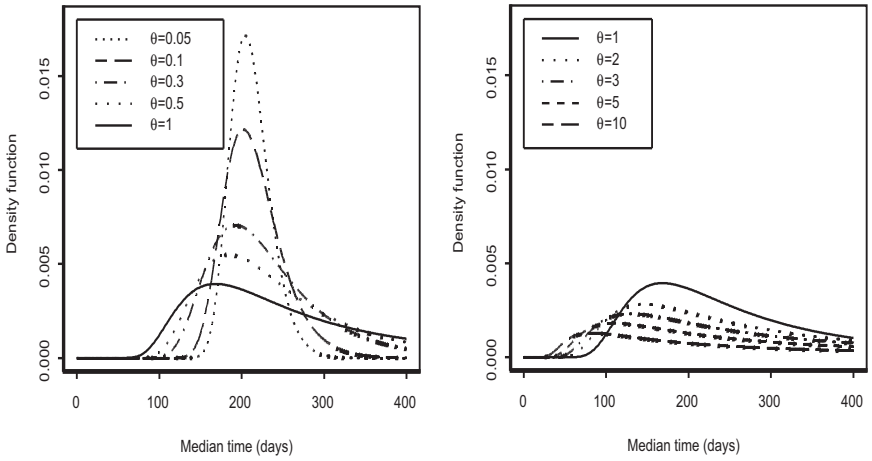

Figure 7.2: Density function for the median time to infection with *Corynebacterium bovis* for different values of θ. Left panel: values of θ between 0 and 1. Right panel: values of θ between 1 and 10.

clustering in the data (represented by the parameter θ) on the median time to infection for the four bacteria is similar.

The interpretation of the covariate effects is different in the marginal and copula model versus the fixed effects and frailty model. Covariate effects in the marginal and copula model are at the marginal level, while covariate effects in the fixed effects and frailty model are at the conditional level. For the location covariate, for example, this means that in the marginal and copula model the hazard ratio represents the ratio of the hazard of infection for a randomly chosen rear udder quarter versus the hazard of infection for a randomly chosen front udder quarter from whatever other cow. In the fixed effects and frailty model the hazard ratio reflects the hazard of infection for a rear versus a front udder quarter within the same cow.

To investigate possible correlation structures in the mastitis data, different correlated gamma frailty models are proposed. The assumption of gamma-distributed frailties makes it possible to obtain a closed form expression for

the marginal likelihood, which can then be maximized to obtain parameter estimates. Variance estimates can be obtained from the Hessian matrix evaluated at the maximum likelihood solutions. If a parametric distribution is assumed for the marginal survival functions, two-stage and one-stage estimation is possible. If the marginal survival functions are estimated nonparametrically or semiparametrically, a two-stage estimation procedure has to be used. In this thesis we restrict to parametric one-stage estimation.
In the proposed correlated frailty models σ^2 is the variance of the frailties of the udder quarters. We first discuss the interpretation of this parameter. In the correlated frailty model σ^2 needs to be interpreted as a measure of the heterogeneity in the data. Since the frailties consist of a part that is common for all or some udder quarters and a part that is specific to a certain udder quarter it can be investigated whether variability is mostly at the cow level or at the udder quarter level. This is not possible in the shared frailty model: all variability is at the cow level.
Different correlated gamma frailty models with different correlation structures are proposed. In the shared gamma frailty model the correlation structure between the frailties is symmetric and the correlation between any frailty-pair is equal to one as all the udder quarters share the same frailty. The first correlated gamma frailty model is a straightforward extension of the correlated gamma frailty model for bivariate data to fourdimensional data. The correlation structure is still symmetric, but the correlation between any frailty-pair is equal to a value ρ, instead of equal to one. Therefore, the correlation between the frailties can be less strong, meaning that there is also considerable variability at the udder quarter level. Between the members of a pair in bivariate data, only one correlation is possible. In fourdimensional data different correlations between frailty-pairs can occur. The other proposed fourdimensional correlated gamma frailty models make an asymmetric and thus more flexible correlation structure between the frailties possible. In certain situations it can be unrealistic to assume that the correlation between all frailties is the same. In the infection with *C. bovis* data set all correlations are very high. Still, a model with a higher correlation between the frailties of the front udder quarters and between the frailties of the rear udder quarters fits the data better than models with a symmetric correlation structure. For other data sets this asymmetry in the correlation structure can be even more pronounced.
Since the correlation parameters represent correlation between frailties and not correlation between infection times, it is hard to interpret them. Yashin et al. (1995) present an approximate formula for the correlation between infection times in terms of the correlation between frailties. Lindeboom and

Van Den Berg (1994) derive explicit expressions for the correlation between the infection times in the special case of a constant baseline hazard function. Unfortunately, such explicit results are not available for more general situations and would therefore make an interesting and necessary topic of further research.

For the data sets and the developed methodology for clustered, interval-censored data considered in this thesis (the copula model (one-stage and two-stage approach), the shared frailty model and the correlated frailty model), using techniques for right-censored data with the midpoint of the interval taken as an exact event time or using the proposed techniques for interval-censored data, provides similar parameter estimates for the covariate effects. Imputation of the upper bound leads to different parameter estimates, especially for the parameter γ. However, simulation studies show that ignoring the interval censoring, and instead for example, using imputation of the midpoint can lead to biased estimates in other situations, for example, in case of a decreasing hazard or when intervals are broad. Therefore, we recommend to use the proposed techniques for interval-censored data instead of midpoint imputation in practice.

The shared frailty model presented in this thesis is limited to one level of clustering. The mastitis data however present two nested levels of clustering: udder quarters are clustered within cow and cows are clustered within herds. It would be interesting to incorporate two levels of clustering in the shared frailty model. This type of model is called a hierarchical model with a frailty term (cow) nested in another frailty term (herd). If more than one frailty term occurs, the frailty terms can no longer be integrated out analytically to obtain a closed form expression for the marginal likelihood. One possible estimation technique is based on numerical integration of the frailties at the higher cluster level. The frailties at the lower level are assumed to be gamma distributed and integrated out analytically (Rondeau et al., 2006). An application of this model for right-censored data can be found in Duchateau and Janssen (2008), p 277. It would be interesting to implement this technique for the interval-censored times to infection in the mastitis data. Another estimation approach is based on Bayesian methodology and uses Gibbs sampling. The normal distribution is a common choice for the distribution of the random effects in hierarchical models; complex hierarchical structures can be easily described by the multivariate normal distribution. It is furthermore easy to fit the models for interval-censored data in a Bayesian context. However, Bayesian methods are not considered

in this thesis. The impossibility to incorporate more than one level of clustering is a bottleneck for the frailty model and the copula model presented in this thesis. The problem could be addressed by including management factors of the herd as fixed effects in the model. That way, parameter estimates are corrected for the included factors, but of course not all factors can be measured or included. Incorporating a second random herd effect in the models would mean a great step forward for these models and is definitely a subject of further research. Determining which level of clustering is more important (clustering within cow or clustering within herd) is a difficult discussion. Interpretation of the results from an analysis with one level of clustering should be careful and a full risk factor analysis of the mastitis data should be postponed until a second level of clustering can be incorporated in the models. Nevertheless, the developed methodology in this thesis already addresses the commonly encountered problems of interval censoring and (one level of) clustering in the data and is therefore a step forward in exploiting all the information available in data sets such as the mastitis data.

In this thesis a Weibull distribution is chosen for the marginal or conditional baseline hazard. Choosing a parametric distribution for the baseline hazard enables us to use standard maximum likelihood techniques in all models. Different distributional assumptions are investigated in the shared gamma frailty model; a Weibull distribution provides the best fit for the mastitis data. A Weibull distribution is therefore assumed for all the models considered in this thesis to make comparison between the models possible. For the infection with *C. bovis* data set for example, the conclusions concerning the effect of the covariates is the same in all proposed models, therefore, the choice for a specific model can be based on the specific interest of the researcher. Usually, the choice of a specific parametric distribution for the baseline hazard is not dictated by biological reasoning. Therefore, a semiparametric approach is sometimes preferred. The expression of the joint survival function in terms of the marginal survival functions in the copula model and the correlated frailty models allows a semiparametric two-stage estimation approach. It would be interesting to investigate the possibility to use the EM-algorithm for semiparametric estimation in the shared gamma frailty model for interval-censored data and in the fourdimensional correlated gamma frailty models for right-censored or interval-censored data, thus using this more flexible approach with less assumptions to model interval-censored, clustered udder quarter infection times.

Bibliography

Aalen, O. O. (1978), "Nonparametric inference for a family of counting processes," *Annals of Statistics*, 6, 701–726.

Abbring, J. H., and Van Den Berg, G. J. (2007), "The unobserved heterogeneity distribution in duration analysis," *Biometrika*, 94, 87–99.

Adkinson, R. W., Ingawa, K. H., Blouin, D. C., and Nickerson, S. C. (1993), "Distribution of clinical mastitis among quarters of the bovine udder," *Journal of Dairy Science*, 76, 3453–3459.

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

Barkema, H. W., Schukken, Y. H., Lam, T. J. G. M., Galligan, D. T., Beiboer, M. L., and Brand, A. (1997), "Estimation of interdependence among quarters of the bovine udder with subclinical mastitis and implications on the analysis," *Journal of Dairy Science*, 80, 1592–1599.

Barrio, B., Vangroenweghe, F., Dosogne, H., and Burvenich, C. (2000), "Decreased neutrophil bactericidal activity during phagocytosis of a slime-producing *Staphylococcus aureus* strain," *Veterinary Research*, 31, 603–609.

Beard, R. E. (1959), "Note on some mathematical mortality models," in *CIBA Foundation Colloquia on Ageing, The Life Span of Animals*, vol. 5, pp. 302–311.

Bebchuk, J. D., and Betensky, R. A. (2000), "Multiple imputation for simple estimation of the hazard function based on interval-censored data," *Statistics in Medicine*, 19, 405–419.

Bellamy, S., Li, Y., Ryan, L., Lipsitz, L., Canner, M., and Wright, R. (2004), "Analysis of clustered and interval-censored data from a community-based study in asthma," *Statistics in Medicine*, 23, 3607–3621.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002), "A local likelihood proportional hazards model for interval-censored data," *Statistics in Medicine*, 21, 263–275.

Bogaerts, K., Leroy, R., Lesaffre, E., and Declerck, D. (2002), "Modelling tooth emergence data based on multivariate interval-censored data," *Statistics in Medicine*, 21, 3775–3787.

Box, G. (1976), "Science and statistics," *Journal of the American Statistical Association*, 71, 791–802.

Breslow, N. E. (1974), "Covariance analysis of censored survival data," *Biometrics*, 30, 88–99.

Burvenich, C., Bannerman, D. D., Lippolis, J. D., Peelman, L., Nonnecke, B., Kehrli, M. E. J., and Paape, M. J. (2007), "Cumulative physiological events influence the inflammatory response of the bovine udder to *Escherichia coli* infections during the transition period," *Journal of Dairy Science*, 90, E39–E50.

Burvenich, C., Monfardini, E., Mehrzad, J., Capuco, A. V., and J, P. M. (2004), "Role of neutrophil polymorphonuclear leukocytes during bovine coliform mastitis: physiology or pathology," *Verhandelingen - Koninklijke Academie voor Geneeskunde van België*, 66, 97–150.

Burvenich, C., Van Merris, V., Mehrzad, J., Diez-Fraile, A., and L, D. (2003), "Severity of *E. coli* mastitis is mainly determined by cow factors," *Veterinary Research*, 34, 521–564.

Cai, T., and Betensky, R. A. (2003), "Hazard regression for interval-censored data with penalized spline," *Biometrics*, 59, 570–579.

Clayton, D., and Cuzick, J. (1985), "Multivariate generalizations of the proportional hazards model," *Journal of the Royal Statistical Society A*, 148, 82–117.

Clayton, D. G. (1978), "A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence," *Biometrika*, 65, 141–151.

Costigan, T. M., and Klein, J. P. (1993), "Multivariate survival analysis based on frailty models," in *Advances in reliability*, pp. 43–58.

Cox, D. R. (1972), "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society, Series B: Methodological*, 34, 187–220.

— (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

Cui, S. F., and Sun, Y. Q. (2004), "Checking for the gamma frailty distribution under the marginal proportional hazards frailty model," *Statistica Sinica*, 14, 249–267.

De Gruttola, V., and Lagakos, S. W. (1989), "Analysis of doubly-censored survival data, with application to AIDS," *Biometrics*, 45, 1–12.

De Vliegher, S., Barkema, H. W., Opsomer, G., de Kruif, A., and Duchateau, L. (2005), "Association between somatic cell count in early lactation and culling of dairy heifers using Cox frailty models," *Journal of Dairy Science*, 88, 560–568.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B: Methodological*, 39, 1–38.

Dosogne, H., Vangroenweghe, F., Barrio, B., Rainard, P., and Burvenich, C. (2001), "Decreased number and bactericidal activity against *Staphylococcus areus* of the resident cells in milk of dairy cows during early lactation," *Journal of Dairy Science*, 68, 539–549.

Duchateau, L., and Janssen, P. (2005), "Understanding heterogeneity in generalized mixed and frailty models," *The American Statistician*, 23, 3607–3621.

— (2008), *The frailty model*, New York: Springer.

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

Dufour, S., Fréchette, A., Barkema, H. W., Mussell, A., and Scholl, D. T. (2011), "Invited review: Effect of udder health management practices on herd somatic cell count," *Journal of Dairy Science*, 94, 563–579.

Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000), "Which copula is the right one?" *Technical report. Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.*

Economou, P., and Caroni, C. (2005), "Graphical tests for the assumption of gamma and inverse Gaussian frailty distributions." *Lifetime Data Analysis*, 11, 565–582.

Fine, J. P., Glidden, D. V., and Lee, K. E. (2003), "A simple estimator for a shared frailty regression model," *Journal of the Royal Statistical Society Series B*, 65, 317–329.

Finkelstein, D. M. (1986), "A proportional hazards model for interval-censored failure time data," *Biometrics*, 42, 845–854.

Genest, C., and MacKay, J. (1986), "The joy of copulas: bivariate distributions with uniform marginals," *American Statistician*, 40, 280–283.

Giard, N., Lichtenstein, P., and Yashin, A. I. (2002), "A multistate model for the genetic analysis of the ageing process," *Statistics in Medicine*, 21, 2511–2526.

Gill, R. D. (1984), "Understanding Cox's regression model: a martingale approach," *Journal of the American Statistical Association*, 79, 441–447.

Glidden, D. V. (1999), "Checking the adequacy of the gamma frailty model for multivariate failure times." *Biometrika*, 86, 381–393.

Goetghebeur, E., and Ryan, L. (2000), "Semiparametric regression analysis of interval-censored data," *Biometrics*, 56, 1139–1144.

Goggins, W. B., Finkelstein, D. M., Schoenfield, D. A., and Zaslavsky, A. M. (1998), "A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model," *Biometrics*, 54, 1498–1507.

Gomez, G., Calle, M. L., and Oller, R. (2004), "Frequentist and Bayesian approaches for interval-censored data," *Statistical Papers*, 2, 139–173.

Griffin, J. (2005), "Intcens: Stata module to perform interval-censored survival analysis," *Boston College Working Papers in Economics*.

Groeneboom, P., and Wellner, J. A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, Boston: Birkhäuser–Verlag.

Harmon, R. J., Eberhart, R. J., Jasper, D. E., Langlois, B. E., and Wilson, R. A. (1990), "Microbiological procedures for the diagnosis of bovine udder infection," *National Mastitis Council, Arlington, VA*.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and coarse data," *The Annals of Statistics*, 19, 2244–2253.

Heyneman, R., Burvenich, C., and Vercauteren, R. (1990), "Interaction between the respiratory burst activity of neutrophil leukocytes and experimentally induced *Escherichia coli* mastitis in cows," *Journal of Dairy Science*, 73, 985–994.

Hoeben, D., Burvenich, C., Eppard, P. J., and Hard, D. L. (1999), "Effect of recombinant bovine somatotropin on milk production and composition

of cows with *Streptococcus uberis* mastitis," *Journal of Dairy Science*, 82, 1671–1683.

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

— (1986b), "Survival models for heterogeneous populations derived from stable distributions," *Biometrika*, 73, 387–396.

— (1999), "Fundamentals of survival data," *Biometrics*, 55, 13–22.

— (2000), *Analysis of multivariate survival data*, New York: Springer.

Iachine, I. A. (1995), "Parameter estimation in the bivariate correlated frailty model with observed covariates via the EM-algorithm." Technical Report Population Studies of Ageing 16, Odense University.

Izumi, S., and Ohtaki, N. (2004), "Aspects of the Armitage-Doll gamma frailty model for cancer incidence data," *Environmetrics*, 15, 209–218.

Jain, N. C. (1979), "Common mammary pathogens and factors in infection and mastitis," *Journal of Dairy Science*, 62, 128–134.

Jongbloed, G. (1998), "The iterative convex minorant algorithm for non-parametric estimation," *Journal of Computational and Graphical Statistics*, 7, 310–321.

Jonker, M. A., and Boomsma, D. I. (2010), "A frailty model for (interval) censored family survival data, applied to the age at onset of non-physical problems," *Lifetime Data Analysis*, 16, 299–315.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Kehrli, M. E. J., Nonnecke, B. J., and Roth, J. A. (1989), "Alterations in bovine neutrophil function during the periparturient period," *American Journal of Veterinary Research*, 50, 207–214.

Kelly, P. J., and Lim, L. L. (2000), "Survival analysis for recurrent event data: an application to childhood infectious diseases," *Statistics in Medicine*, 19, 13–33.

Klein, J. P. (1992), "Semiparametric estimation of random effects using the cox model based on the EM algorithm," *Biometrics*, 48, 795–806.

Klein, J. P., and Moeschberger, M. L. (2003), *Survival analysis: techniques for censored and truncated data*, New York: Springer.

Kooperberg, C., and Stone, C. J. (1992), "Logspline density estimation for censored data," *Journal of Computational and Graphical Statistics*, 1, 301–328.

Laevens, H., Deluyker, H., Schukken, Y. H., De Meulemeester, L., Vandermeersch, R., De Meulenaere, E., and De Kruif, A. (1997), "Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows," *Journal of Dairy Science*, 80, 3219–3226.

Lam, T. J. G. M., Van Vliet, J. H., Schukken, Y. H., Grommers, F. J., Van Velden-Russcher, A., Barkema, H. W., and Brand, A. (1997), "The effect of discontinuation of postmilking teat disinfection in low somatic cell count herds. I. incidence of clinical mastitis," *Veterinary Quarterly*, 19, 41–47.

Lanternier, B., Lyonnet, P., and Toscano, R. (2008), "Mechanical component reliability, hazard proportional model," *Mecanique & Industries*, 9, 397–405.

Lawless, J. F. (2003), *Statistical models and methods for lifetime data*, London: John Wiley.

Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R., and Duchateau, L. (2005), "A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model," *Statistics in Medicine*, 24, 3789–3804.

Lindeboom, M., and Van Den Berg, G. J. (1994), "Heterogeneity in models for bivariate survival: the importance of the mixing distribution," *Journal of the Royal Statistical Society B*, 56, 49–60.

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

Lohuis, J. A., Schukken, Y. H., Henricks, P. A., Heyneman, R., Burvenich, C., Verheijden, J. H., Van Miert, A. S., and Brand, A. (1990), "Preinfection functions of blood polymorphonuclear leukocytes and the outcome of experimental *Escherichia coli* mastitis in the cow," *Journal of Dairy Science*, 73, 342–350.

Manatunga, A. K., and Oakes, D. (1999), "Parametric analysis for matched pair survival data," *Lifetime Data Aalysis*, 5, 371–387.

Martinussen, T., and Pipper, C. B. (2005), "Estimation in the positive stable shared frailty Cox proportional hazards model," *Lifetime data analysis*, 11, 99–115.

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.

McGilchrist, C. A. (1993), "REML estimation for survival models with frailty," *Biometrics*, 49, 221–225.

McGilchrist, C. A., and Aisbett, C. W. (1991), "Regression with frailty in survival analysis," *Biometrics*, 47, 461–466.

Mehrzad, J., Duchateau, L., Pyorala, S., and Burvenich, C. (2002), "Blood and milk neutrophil chemiluminescence and viability in primiparous and pluriparous dairy cows during late pregnancy, around parturition and early lactation," *Journal of Dairy Science*, 85, 3268–3276.

Neijenhuis, F., Barkema, H. W., Hogeveen, H., and Noordhuizen, J. P. T. M. (2001), "Relationship between teat end callosity and occurrence of clinical mastitis," *Journal of Dairy Science*, 84, 2664–2672.

Nelsen, R. (1996), "Nonparametric measures of multivariate association," *Lecture Notes-Monograph Series*, 28, 223–232.

Nelsen, R. B. (2006), *An introduction to copulas*, New York: Springer-Verlag.

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

Nickell, S. (1979), "Estimating the probability of leaving unemployment," *Econometrica*, 47, 1249–1266.

Oakes, D. (1982), "A model for association in bivariate survival data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 44, 414–422.

— (1989), "Bivariate survival models induced by frailties," *Journal of the American Statistical Association*, 84, 487–493.

O'Brien, B., Fitzpatrick, C., Meaney, W. J., and Joyce, P. (1999), "Relationship between somatic cell count and neutrophils in milk," *Irish Journal of Agricultural and Food Research*, 38, 288–296.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992), "Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model," *Biometrics*, 48, 951–959.

Oehlert, G. W. (1992), "A note on the delta method," *The American Statistician*, 46, 27–29.

Oliver, S. P., and Sordillo, L. M. (1988), "Udder health in the periparturient period," *Journal of Dairy Science*, 71, 2584–2606.

Oller, R., Gomez, G., and Calle, M. L. (2004), "Interval censoring: model characterizations for the validity of the simplified likelihood," *The Canadian Journal of Statistics*, 32, 315–326.

Orbe, J., Ferreira, E., and Nunez-Anton, V. (2002), "Comparing proportional hazards and accelerated failure time models for survival analysis," *Statistics in Medicine*, 21, 3493–3510.

Pan, W. (2000), "A multiple imputation approach to Cox regression with interval-censored data," *Biometrics*, 56, 199–203.

Peto, R. (1973), "Experimental survival curves for interval-censored data," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 22, 86–91.

Ripatti, S., and Palmgren, J. (2000), "Estimation of multivariate frailty models using penalised partial likelihood," *Biometrics*, 56, 1016–1022.

Risselada, M., Kramer, M., De Rooster, H., Taeymans, O., Verleyen, P., and Van Bree, H. (2005), "Ultrasonographic and radiographic assessment of uncomplicated secondary fracture healing of long bones in dogs and cats," *Veterinary Surgery*, 34, 99–107.

Rondeau, V., Commenges, D., and Joly, P. (2003), "Maximum penalized likelihood estimation in a gamma frailty model," *Lifetime Data Analysis*, 9, 139–153.

Rondeau, V., Filleul, L., and Joly, P. (2006), "Nested frailty models using maximum penalized likelihood estimation," *Statistics in Medicine*, 25, 4036–4052.

Royall, R. M. (1986), "Model robust confidence intervals using maximum likelihood estimators," *International Statistical Review*, 54, 221–226.

Satten, G. A. (1996), "Rank-based inference in the proportional hazards model for interval-censored data," *Biometrika*, 83, 355–370.

Satten, G. A., Datta, S., and Williamson, J. M. (1998), "Inference based on imputed failure times for the proportional hazards model with interval-censored data," *Journal of the American Statistical Association*, 93, 318–327.

Schepers, A. J., Lam, T. J. G. M., Schukken, Y. H., Wilmink, J. B. M., and Hanekamp, W. J. A. (1997), "Estimation of variance components for somatic cell counts to determine thresholds for uninfected quarters," *Journal of Dairy Science*, 80, 1833–1840.

Schukken, Y. H., Leslie, K. E., Barnum, D. A., Mallard, B. A., Lumsden, J. H., Dick, P. C., Vessie, G. H., and Kehrli, M. E. (1999), "Experimental *Staphylococcus aureus* intramammary challenge in late lactation dairy cows: quarter and cow effects determining the probability of infection," *Journal of Dairy Science*, 82, 2393–2401.

Sears, P. M., Smith, B. S., English, P. B., Herer, P. S., and Gonzalez, R. N. (1990), "Shedding pattern of *Staphylococcus aureus* from bovine intramammary infections," *Journal of Dairy Science*, 73, 2785–2789.

Seegers, H., Fourichon, C., and Beaudeau, F. (2003), "Production effects related to mastitis and mastitis economics in dairy cattle herds," *Journal of Animal Ecology*, 72, 640–649.

Self, S. G., and Grossman, E. A. (1986), "Linear rank tests for interval-censored data with applications to PCB levels in adipose tissue of transformer repair workers," *Biometrics*, 42, 521–530.

Shih, J. H., and Louis, T. A. (1995a), "Assessing gamma frailty models for clustered failure time data," *Lifetime Data Analysis*, 1, 205–220.

— (1995b), "Inferences on the association parameter in copula models for bivariate survival data," *Biometrics*, 51, 1384–1399.

Sklar, A. (1959), "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'institut de statistique de l'université de Paris*, 8, 229–231.

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.

StataCorp. (2005), *Stata Statistical Software: Release 9*, College Station.

Sun, J. (2006), *The statistical analysis of interval-censored failure time data*, New York: Springer.

Sun, L., Kim, Y., and Sun, J. (2004), "Regression analysis of doubly censored failure time data using the additive hazards model," *Biometrics*, 60, 637–643.

Sun, L., Wang, L., and Sun, J. (2006), "Estimation of the association for bivariate interval-censored failure time data," *Scandinavian Journal of Statistics*, 33, 637–649.

Tanner, M. A., and Wong, W. H. (1987), "The application of imputation to an estimation problem in grouped lifetime analysis," *Technometrics*, 29, 23–32.

Therneau, T. M., and Grambsch, P. M. (2000), *Modelling survival data: extending the Cox model*, New York: Springer Verlag.

Turnbull, B. W. (1976), "The empirical distribution with arbitrarily grouped censored and truncated data," *Journal of the Royal Statistical Society, Series B: Methodological*, 38, 290–295.

Turnbull, B. W., and Weiss, L. (1978), "A likelihood ratio statistic for testing goodness of fit with randomly censored data," *Biometrics*, 34, 367–375.

Vangroenweghe, F., Rainard, P., Paape, M., Duchateau, L., and Burvenich, C. (2004), "Increase of *Escherichia coli* inoculum doses induce faster innate immune respons in primiparous cows," *Journal of Dairy Science*, 87, 4132–4144.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, 16, 439–454.

Vecht, U., Wisselink, H. J., and Defize, P. R. (1989), "Dutch national mastitis survey. the effect of herd and animal factors on SCC," *Netherlands Milk and Dairy Journal*, 43, 425.

Viswanathan, B., and Manatunga, A. K. (2001), "Diagnostic plots for assessing the frailty distribution in multivariate survival data," *Lifetime Data Analysis*, 7, 143–155.

Wang, S. T., Klein, J. P., and Moeschberger, M. L. (1995), "Semiparametric estimation of covariate effects using the positive stable frailty model," *Applied Stochastic Models and Data Analysis*, 11, 121–133.

Wei, L. J., and Glidden, D. V. (1997), "An overview of statistical methods for multiple failure time data in clinical trials," *Statistics in Medicine*, 16, 833–839.

Weibull, W. (1951), "A statistical distribution function of wide applicability," *Journal of Applied Mechanics-Transactions of the ASME*, 18, 293–297.

Weller, J. I., Saran, A., and Zeliger, Y. (1992), "Genetic and environmental relationships among somatic cell count, bacterial infection, and clinical mastitis," *Journal of Dairy Science*, 75, 2532–2540.

Wellner, J. A., and Zhan, Y. (1997), "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data," *Journal of the American Statistical Association*, 92, 945–959.

Wienke, A., Arbeev, K. G., Locatelli, I., and Yashin, A. I. (2005a), "A comparison of different bivariate correlated frailty models and estimation strategies," *Mathematical Biosciences*, 198, 1–13.

Wienke, A., Christensen, K., Skytthe, A., and Yashin, A. I. (2002), "Genetic analysis of cause of death in a mixture model of bivariate lifetime data," *Statistical Modelling*, 2, 89–102.

Wienke, A., Herskind, A. M., Christensen, K., Skytthe, A., and Yashin, A. I. (2005b), "The heritability of CHD mortality in Danish twins after controlling for smoking and BMI," *Twin Research and Human Genetics*, 8, 53–59.

Wienke, A., Ripatti, S., Palmgren, J., and Yashin, A. (2010), "A bivariate survival model with compound Poisson frailty," *Statistics in Medicine*, 29, 275–283.

Wu, C. F. J. (1986), "Jackknife, bootstrap and other resampling methods in regression analysis," *The Annals of Statistics*, 33, 637–649.

Xue, X., and Brookmeyer, R. (1996), "Bivariate frailty model for the analysis of multivariate survival time," *Lifetime Data Analysis*, 2, 277–289.

Yashin, A. I., and Iachine, I. A. (1997), "How frailty models can be used for evaluating longevity limits: taking advantage of an interdisciplinary approach," *Demography*, 34, 31–48.

— (1999), "Dependent hazards in multivariate survival problems," *Journal of Multivariate Analysis*, 71, 241–261.

Yashin, A. I., Vaupel, J. W., and Iachine, I. A. (1995), "Correlated individual frailty: an advantageous approach to survival analysis of bivariate data," *Mathematical Population Studies*, 5, 145–159.

Yau, K., and McGilchrist, C. (1997), "Use of generalised linear mixed models for the analysis of clustered survival data," *Biometrical Journal*, 39, 3–11.

Zadoks, R. N., Allore, H. G., Barkema, H. W., Sampimon, O. C., Wellenberg, G. J., Gröhn, Y. T., and Schukken, Y. H. (2001), "Cow- and quarter-level risk factors for *Streptococcus uberis* and *Staphylococcus aureus* mastitis," *Journal of Dairy Science*, 81, 2649–2663.

Zdravkovic, S., Wienke, A., Pedersen, N. L., Marenberg, M. E., Yashin, A. I., and de Faire, U. (2004), "Genetic influences on CHD-death and the impact of known risk factors: comparison of two frailty models," *Behavior Genetics*, 34, 585–592.

Summary

The mastitis data set and its complex structure is the source of inspiration for this thesis. Standard survival analysis techniques assume that the population under study is homogeneous and that the event times are independent. Furthermore, the most common and widely used survival analysis models are developed for right-censored data. The mastitis data set, however, has two characteristics that require extension of the currently available survival analysis techniques if they have to be dealt with simultaneously. First, the data are hierarchically structured, with observational units (the udder quarters) grouped in blocks (the cow), so that the event times within a cow can not be assumed to be independent. Second, since the udder quarters are sampled only (more or less) monthly, the time to infection is not known exactly; it is only known that the infection happened between the last visit with a negative test and the first visit with a positive test; therefore, the infection time is interval-censored.

We first compare the existing methodologies to model hierarchical, interval-censored event time data. We first consider the modeling of hierarchical data, handling the interval censoring problem by imputing the midpoint of the interval as an exact event time. The frailty model and the copula model are two models that can be used to fit correlated, right-censored event time data. Both models provide an estimate of the correlation between event times in a cluster. In Chapter 3 similarities and differences between copula models and shared frailty models are discussed. Focus is on the comparison between the Clayton-Oakes copula model and the shared gamma frailty model; and between the positive stable copula model and the shared positive stable frailty model. The comparison reveals that the correlation structure used to obtain the joint survival function from the marginal survival functions in both models is the same, but that the arguments in the joint survival function, the marginal survival functions, are modeled in a different way. The marginal survival functions in the copula model are assumed to be Weibull distributed, but the marginal survival functions in the frailty model are obtained by integrating out the frailty from the conditional survival functions (assumed to be Weibull distributed) and contain the parameter θ, contrary to the marginal survival functions in the copula model. The differences are shown by using the Clayton-Oakes copula model with Weibull marginal survival functions and the shared gamma frailty model with conditional Weibull survival functions. A similar comparison between the positive stable copula model and the shared positive stable frailty model shows that, in the exclusive case of a Weibull baseline hazard (with $\gamma \neq 1$), there is a one-to-one match between the two models. If, for example, the exponential distribution (Weibull distribution with $\gamma = 1$) is assumed for

the event times together with a positive stable distribution for the frailties, this property no longer holds.

In Chapter 4 the interval-censored nature and the hierarchical structure is dealt with simultaneously. The most frequently used methods to analyze clustered, interval-censored data, available in commercial software packages, are discussed: the marginal model, the fixed effects model and the copula model. We point out some disadvantages or shortcomings of these models, in particular related to the mastitis data. The marginal model and the fixed effects model do not provide an estimate of the correlation and should only be used if interest is restricted to the covariate effects. Since, in the marginal model, the existing correlation between event times is ignored when obtaining parameter estimates, the likelihood-based estimates of the variance of the estimates are not consistent. Therefore, other techniques, such as the grouped jackknife technique, should be used to obtain estimates of the variance. In the fixed effects model the cluster effect is modeled by adding a fixed effect for each cluster. However, caution is needed when applying a fixed effects model to the mastitis data. Cow level covariates can not be modeled in the fixed effects model because there is complete confounding between the fixed effect factors for cow and the cow level covariates. In the copula model, a two-stage estimation approach is necessary if one wants to model the interval censoring in the data. In the first stage, a marginal model for interval-censored data can be fitted, but in the second stage the only option is to maximize the likelihood for right-censored data, using midpoint imputation for interval-censored observations, since the likelihood for interval-censored data is not available in the literature. Since it would be interesting to model the interval censoring in both stages of the two-stage estimation procedure in the copula model, we further describe the construction of the likelihood for fourdimensional interval-censored data in the copula model so that interval censoring is used throughout. Making a parametric assumption for the baseline hazard allows, next to the two-stage procedure, a one-stage estimation approach. Herewith we introduce a new approach to model clustered, interval-censored data.

In Chapter 5 an extension of the parametric shared gamma frailty model to interval-censored data is proposed. We show that a closed form expression of the marginal likelihood can be obtained by integrating out the gamma-distributed frailties, which can then be maximized to obtain parameter estimates. Furthermore, second derivatives of the likelihood can be derived and thus estimates for the variances of the parameters can be obtained by inverting the matrix of second derivatives. The technique allows the inclusion of covariates in the model and is characterized by little or no data constraints.

Contrary to the one-stage copula model discussed in Chapter 4 the number of cluster members can be variable. Intervals can be of variable length, though the parameter γ tends to be more and more biased when intervals become broader. A simulation study shows that the proposed technique outperforms imputation techniques, especially in the case of a decreasing Weibull baseline hazard or when censoring intervals get broader.
To investigate the correlation structure between the udder quarters fourdimensional correlated gamma frailty models are proposed in Chapter 6. The proposed models allow different correlation structures between the hazards of the udder quarters (and thus also between the event times). The correlation between the hazards of two udder quarters is modeled by imposing a correlation structure on the two frailties. First, the symmetric correlated frailty model (with the correlation equal to ρ for all frailty pairs) is described for fourdimensional data. Next, we define a model that is intermediate between the shared gamma frailty model and the symmetric correlated frailty model. The correlation between the frailties of the front udder quarters and between the frailties of the rear udder quarters is equal to one, while the correlation between other frailty pairs is equal to ρ. The final model has the most flexible correlation structure with the correlation between the frailties of the front udder quarters and between the frailties of the rear udder quarters equal to ρ_1 and the correlation between other frailty pairs equal to ρ_2. By assuming a gamma distribution for the frailties, the frailties can be integrated out from the conditional joint survival function and a closed form expression for the marginal likelihood based on the derivatives of the fourdimensional joint survival function is obtained which can then be maximized to obtain parameter estimates and their standard errors. The representation of the joint survival function in terms of the marginal survival functions and the correlation parameters allows a semiparametric two-stage estimation approach. The models provide insight in the most likely correlation structure, and in the strength of the correlation between the frailties of the udder quarters, clustered in the cow. Each of the correlated frailty models described above imposes certain constraints on the correlation between the frailties, e.g. $\rho_2 < \rho_1$, meaning that the correlation between the front or rear udder quarters is necessarily stronger that the correlation between the left or right udder quarters. However, in the context of the mastitis data set such constraints seem to be realistic.

Samenvatting

Deze thesis werd geïnspireerd door een data set met gegevens rond bacteriële infecties op het niveau van een uierkwartier bij de melkkoe, de mastitis data set. Standaard technieken in de overlevingsanalyse veronderstellen dat de bestudeerde populatie homogeen is en dat de infectietijden onafhankelijk zijn. Bovendien zijn de bekendste technieken ontwikkeld voor rechts gecensureerde gegevens. Twee eigenschappen van de mastitis data set maken het echter noodzakelijk de beschikbare technieken in de overlevingsanalyse uit te breiden. Vooreerst zit er een bepaalde hiërarchie in de data, de uierkwartieren zijn gegroepeerd binnen een koe, waardoor we niet kunnen aannemen dat de infectietijden binnen een koe onafhankelijk zijn. Verder is de exacte infectietijd niet gekend gezien de uierkwartieren slechts ongeveer maandelijks bemonsterd worden; zodoende weten we enkel dat de infectie plaatsvond tussen het laatste bezoek met een negatieve test en het eerste bezoek met een positieve test. De infectietijd is daardoor interval gecensureerd.
We vergelijken eerst een aantal bestaande technieken die gebruikt worden om hiërarchische, interval gecensureerde gegevens te modelleren. We bespreken eerst het modelleren van hiërarchische gegevens waarbij de interval censurering aangepakt wordt door het midden van het interval te gebruiken als een exact gekende infectietijd. Het frailty model en het copula model worden vaak aangewend om gecorreleerde, rechts gecensureerde overlevingsgegevens te modelleren. Beide modellen geven een schatting van de correlatie tussen de infectietijden in een groep. In hoofdstuk 3 bespreken we gelijkenissen en verschillen tussen copula modellen en shared frailty modellen. We focussen vooral op de vergelijking tussen het Clayton-Oakes copula model en het shared gamma frailty model; en tussen het positive stable copula model en het positive stable frailty model. Uit de vergelijking blijkt dat de correlatie structuur die gebruikt wordt om de gemeenschappelijke overlevingsfunctie in functie van de marginale overlevingsfuncties te beschrijven, hetzelfde is in beide modellen, maar dat de argumenten van de gemeenschappelijke overlevingsfunctie, de marginale overlevingsfuncties, in beide modellen anders gemodelleerd worden. De marginale overlevingsfuncties in het copula model volgen een Weibull distributie, maar de marginale overlevingsfuncties in het frailty model bekomt men door de frailty uit te integreren uit de conditionele overlevingsfuncties (die een Weibull distributie volgen) waardoor ze de parameter θ bevatten, integenstelling tot de marginale overlevingsfuncties in het copula model. De verschillen worden aangetoond aan de hand van het Clayton-Oakes copula model met Weibull marginale overlevingsfuncties en aan de hand van het shared gamma frailty model met conditionele Weibull overlevingsfuncties. Een gelijkaardige vergelijking tussen het positive stable copula model en het positive stable shared frailty model toont aan dat in

het uitzonderlijke geval van een Weibull (met $\gamma \neq 1$) basis uitvalsfunctie de twee modellen equivalent zijn. Als bijvoorbeeld verondersteld wordt dat de infectietijden exponentieel verdeeld zijn, zijn de twee modellen niet langer equivalent.
In hoofdstuk 4 worden technieken beschreven die de interval censurering en de hiërarchie in de data tegelijkertijd modelleren. Modellen die vaak gebruikt worden om hiërarchische, interval gecensureerde gegevens te analyseren en bovendien beschikbaar zijn in de commerciële software pakketten worden besproken: het marginale model, het fixed effects model en het copula model. We leggen de nadruk op enkele tekortkomingen of nadelen van deze modellen, als we ze willen gebruiken om de mastitis data te analyseren. Het marginale model en het fixed effects model geven ons geen schatting van de correlatie in de data en kunnen dus enkel gebruikt worden wanneer de onderzoeker enkel geïnteresseerd is in het effect van de covariaten. Gezien de bestaande correlatie tussen de infectietijden genegeerd wordt in het marginale model, zijn de variantieschatters niet consistent. Daarom moeten andere technieken, zoals de gegroepeerde jackknife techniek, aangewend worden om correcte variantieschatters te bekomen. In het fixed effects model wordt voor elke cluster een vast effect toegevoegd. Voorzichtigheid is echter geboden als men de mastitis data wil analyseren met een fixed effects model. Het is niet mogelijk het effect van covariaten op koe-niveau te modelleren omdat zij op hetzelfde niveau van de vaste effecten zitten. In het copula model kan enkel in de eerste stap van de schattingsprocedure rekening gehouden worden met de interval censurering. In de tweede stap wordt de aannemelijkheidssfunctie voor rechts gecensureerde data gemaximaliseerd met het midden van het interval als de gekende infectietijd. Het zou echter interessant zijn om rekening te kunnen houden met de interval censurering in de gegevens in beide stappen van de tweestappen schattingsprocedure in het copula model. Daarom beschrijven we in hoofdstuk 4 hoe de aannemelijkheidssfunctie voor vierdimensionale interval gecensureerde gegevens in het copula model kan opgebouwd worden. Het veronderstellen van een parametrische uitvalsfunctie maakt naast de tweestappen procedure ook een één-stap schattingsprocedure mogelijk. Dit is een nieuwe aanpak om geclusterde, interval gecensureerde gegevens te modelleren.
In hoofdstuk 5 stellen we een uitbreiding voor van het parametrische shared gamma frailty model voor interval gecensureerde gegevens. We tonen aan dat we de marginale aannemelijkheidsfunctie in een gesloten vorm kunnen opschrijven door de frailties, die een gamma distributie volgen, uit te integreren. De marginale aannemelijkheidsfunctie kan dan gemaximaliseerd

worden om de parameter estimates te bekomen. Variantieschatters kunnen bekomen worden uit de inverse van de matrix van tweede afgeleiden van de marginale aannemelijkheidsfunctie. Het model kan covariaten bevatten en er zijn weinig of geen voorwaarden waaraan de data moeten voldoen. Het aantal leden van de clusters kan variëren, integenstelling tot in het één-stap copula model besproken in hoofstuk 4. De breedte van de intervallen kan ook variëren, maar de vertekening van de parameter γ wordt wel groter naargelang de intervallen breder worden.
Om de correlatie structuur binnen de uierkwartieren te bestuderen, stellen we in hoofdstuk 6 vierdimensionale gecorreleerde gamma frailty modellen voor. Deze modellen laten verschillende correlaties tussen uitvalsfuncties van de uierkwartieren toe (en daardoor ook tussen de infectietijden). De correlatie tussen de uitvalsfuncties van twee uierkwartieren komt tot stand door een correlatie structuur op te leggen aan de frailties. Eerst beschrijven we het symmetrische gecorreleerde gamma frailty model voor vierdimensionale gegevens. In dit model is de correlatie tussen elk frailty-paar gelijk aan ρ. Vervolgens definiëren we een model dat tussen het shared gamma frailty model en het symmetrische gecorreleerde frailty model ligt. De correlatie tussen de frailties van de voorste uierkwartieren en tussen de frailties van de achterste uierkwartieren is 1 terwijl de correlatie tussen andere frailty-paren gelijk is aan ρ. Het laatste model heeft de meest flexibele correlatie structuur. De correlatie tussen de frailties van de voorste uierkwartieren en tussen de frailties van de achterste uierkwartieren is niet langer noodzakelijk gelijk aan 1, maar aan een waarde ρ_1. De correlatie tussen andere frailty-paren is gelijk aan ρ_2. Door een gamma distributie te veronderstellen voor de frailties, kunnen deze uitgeïntegreerd worden uit de conditionele overlevingsfunctie. Vervolgens kan een gesloten vorm bekomen worden voor de marginale aannemelijkheidsfunctie gebaseerd op de afgeleiden van de vierdimensionale gemeenschappelijke overlevingsfunctie, die dan gemaximaliseerd kan worden. Doordat de gemeenschappelijke overlevingsfunctie wordt uitgedrukt in functie van de marginale overlevingsfuncties, is een semiparametrische twee-stappen schattingsprocedure ook mogelijk. Deze modellen geven ons inzicht in de meest waarschijnlijke correlatiestructuur en in de sterkte van de correlatie tussen de frailties van de uierkwartieren. Elk van de beschreven gecorreleerde frailty modellen legt bepaalde beperkingen op aan de correlatie tussen de frailties. Zo is bv, $\rho_2 < \rho_1$ in het laatst beschreven model, waardoor de correlatie tussen de voor- of achterkwartieren noodzakelijk sterker is dan de correlatie tussen de linker- en rechterkwartieren. In de context van de mastitis data set lijken deze beperkingen echter realistisch.

Dankwoord

Als ik de blikken zou kunnen registreren die ik toegeworpen krijg als ik vertel dat ik doctoreer in de statistiek, zou dit ongetwijfeld een komische film opleveren. En toch was het statistiek en niet wiskundige natuurkunde of sterrenkunde die me na mijn licentiaatstudie wist te boeien. Statistische analyses zijn onmisbaar in wetenschappelijk onderzoek en als statisticus hoop ik dan ook nog lang een bijdrage te kunnen leveren aan het wetenschappelijk onderzoek.
Dit doctoraatsonderzoek is tot stand gekomen dankzij de hulp en steun van vele mensen. Ondanks het risico mensen te vergeten, wil ik toch een poging doen enkele mensen speciaal te bedanken.
Vooreerst en in het bijzonder wil ik mijn promotor Prof. Dr. Luc Duchateau bedanken omdat hij mij niet alleen de kans gegeven heeft aan mijn doctoraatstudie te beginnen, maar vooral omdat hij mij ook de gelegenheid gegeven heeft om ze succesvol af te ronden. Zijn uitvoerige kennis over het frailty model vormde de basis voor de nieuwe ideeën uitgewerkt in deze thesis. Luc, hartelijk dank voor uw kritische blik en leerrijke adviezen. Dank zij u heb ik deze thesis tot een goed einde kunnen brengen.
Mijn tweede woord van dank gaat uit naar mijn copromotor Prof. Dr. Paul Janssen. Zijn rust en gestructureerdheid brachten geregeld orde in de chaos. Paul, bedankt voor de opbouwende kritiek en de aangename samenwerking.
Een speciaal woord van dank gaat uit naar mijn goede collega Bart Ampe. Hoe vaak hebben we niet gezegd dat een dierenarts en een wiskundige de ideale combinatie vormt in een onderzoeksgroep biometrie aan de faculteit diergeneeskunde. Met mijn diergeneeskundige vragen kon ik altijd bij jou terecht, maar de vele discussies rond statistische problemen hebben zeker ook bijgedragen aan dit doctoraatswerk, om nog maar te zwijgen van je computerkennis. Ik ben blij te mogen zeggen dat ik je niet enkel als een goede collega beschouw, maar vooral als een goede vriend op wie ik altijd kan rekenen. Bart, hartelijk dank voor alles wat je voor me gedaan hebt.
Verder ook een woord van dank voor alle collega's van de vakgroep Vergelijkende fysiologie en biometrie, een kleine maar vooral fijne vakgroep. De frisse wind die afgelopen jaar door onze vakgroep is komen waaien, heeft alvast bij mij het enthousiasme voor wetenschappelijk onderzoek weer aangewakkerd. Ik wens jullie allen een mooie toekomst toe.
Bedankt ook aan het Bijzonder Onderzoeksfonds van de Universiteit Gent en het Interuniversity Attraction Poles (IAP) research network die dit onderzoek financieel ondersteund hebben.
Ook familie en vrienden mogen niet ontbreken in dit dankwoord. Al was mijn doctoraat soms wat een te mijden onderwerp, nu ben ik blij en trots dat ik het tot een goed einde gebracht heb. Dit was niet gelukt zonder jullie

steun, interesse en bemoedigende woorden.
And last but not least, wil ik de persoon bedanken die me nauwst aan het hart ligt. Filip, een doctoraat afwerken is geen eenvoudige opdracht, niet voor de doctorandus, maar ook niet voor zijn of haar partner. Frustratie, stress en tijdsdruk zorgden thuis ongetwijfeld geregeld voor een humeurig vrouwtje terwijl momenten van succes misschien eerder en enkel op de werkvloer gevierd werden. Ik wil je bedanken voor je geduld, steun en behulpzaamheid, vooral in de eindspurt naar het indienen en verdedigen van dit werk. Nu heeft het leven een heel andere uitdaging voor ons in petto. Ik ben er helemaal klaar voor.