A D-vine copula based model for repeated measurements extending linear mixed models with homogeneous correlation structure

Matthias Killiches[*][†]   Claudia Czado[*]

May 17, 2017

Abstract

We propose a model for unbalanced longitudinal data, where the univariate margins can be selected arbitrarily and the dependence structure is described with the help of a D-vine copula. We show that our approach is an extremely flexible extension of the widely used linear mixed model if the correlation is homogeneous over the considered individuals. As an alternative to joint maximum-likelihood a sequential estimation approach for the D-vine copula is provided and validated in a simulation study. The model can handle missing values without being forced to discard data. Since conditional distributions are known analytically, we easily make predictions for future events. For model selection we adjust the Bayesian information criterion to our situation. In an application to heart surgery data our model performs clearly better than competing linear mixed models.

Keywords: *Vine copulas, linear mixed models, repeated measurements, longitudinal data, unbalanced setting.*

1   Introduction

Repeated measurements that are obtained in a longitudinal study are common in many areas. Very early applications in astronomy (Airy, 1861) were followed by a vast number of studies in fields such as industry (e.g. Newbold, 1927), ecology (e.g. Potvin et al., 1990), biology (e.g. Yeung et al., 2003), psychology (e.g. Lorch and Myers, 1990), medicine (e.g. Ludbrook, 1994), education (e.g. Malin and Linnakylä, 2001) and many more.

Over the years many concepts have been developed for the analysis of such repeated measurements. An extensive review on the origins of longitudinal data models can be found in Chapter 1 of Fitzmaurice et al. (2008). Davis (2002) offers a thorough

---

[*]Zentrum Mathematik, Technische Universität München, Boltzmannstraße 3, 85748 Garching, Germany

[†]Corresponding author, email: `matthias.killiches@tum.de`.

introduction to the topic, starting with basic aspects of repeated measurement data. Besides foundations and different modeling aspects of repeated measurement data Lindsey (1999) addresses the question how to design a study. Diggle and Donnelly (1989) give an extensive review on different approaches to the analysis of repeated measurements. The most popular model class for this purpose are probably *linear mixed models* (LMMs). They extend classical linear models by adding individual-specific random effects to the fixed effects. Extensive introductions to this topic can be found for example in Diggle (2002) and Verbeke and Molenberghs (2009). Although the covariance structure of linear mixed models can be fitted rather flexibly, the dependence always remains Gaussian by definition.

Within the last two decades dependence modeling has become more and more popular in all areas of applications. Especially copulas have gained large popularity since they allow to model marginal distributions and the dependence structure separately (Sklar, 1959). Consequently, copulas were also applied for modeling repeated measurement data. This approach has first been used by Meester and MacKay (1994) who developed a model for bivariate clustered categorical data. Lambert and Vandenhende (2002) present a model for multivariate repeated measurement data, where the dependence is described by copula (although only the Gaussian copula is used in the application). Shen and Weissfeld (2006) model serial dependence for continuous longitudinal data with a non-ignorable non-monotone missing-data process using a Gaussian copula. Other examples are Lindsey and Lindsey (2006), who use the Gaussian copula among other multivariate models with correlation matrices for non-linear repeated measurements. Further, Sun et al. (2008) argue that elliptical copulas are better suited than Archimedean copulas for modeling serial dependence in the context of longitudinal data.

D-vine copulas are a special class of *vine copulas* (Bedford and Cooke, 2002; Aas et al., 2009) that are particularly suited for modeling serial dependence. Smith et al. (2010) used them to model longitudinal data in a Bayesian approach. Multivariate time series are considered in Smith (2015) and Nai Ruscone and Osmetti (2017). In Joe (2014, Chapter 7.5) discrete longitudinal count data are modeled using D-vines. Shi and Yang (2016) use a mixed D-vine to model semi-continuous longitudinal claims. All these references work in a balanced setting, i.e. each individual has the same number of measurements. An unbalanced setting is considered by Shi et al. (2016) using a Gaussian copula.

The novelty of the approach presented in this paper is that we develop a D-vine copula based model with arbitrary margins for modeling unbalanced longitudinal data with the aim of understanding the underlying relationship among the measurements and enabling predictions for future events. For prediction we use conditional quantiles that are analytically given. For model selection we derive an adjustment of the *Bayesian information criterion* (BIC) for the proposed model. The model will furthermore be shown to be an extension of a very wide class of linear mixed models for which the correlation matrix of the measurements is homogeneous over the individuals.

Section 2 briefly introduces D-vine copulas and the proposed D-vine copula based model for repeated measurement data. Linear mixed models and their connection with the D-vine based model are developed in Section 3. Section 4 contains maximum-likelihood based estimation methods for the D-vine based model. Further, as a tool for model selection, an adjustment of the BIC for the proposed model is derived. The performance of the estimation methods is investigated in a simulation study (Section 5). In Section 6 we fit both linear mixed models and D-vine based models to a heart surgery

data set and compare the results using likelihood based model selection criteria and performing conditional quantile prediction. contains our conclusions and an outlook on future research.

2 D-vine based repeated measurement model

2.1 Setting and marginal modeling

Consider a repeated measurement (longitudinal) data set $\mathcal{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}$ that contains $n \in \mathbb{N}$ observation blocks $\mathbf{y}^i = (y_1^i, \ldots, y_{d_i}^i)^\top \in \mathbb{R}^{d_i}$ associated with individual i having $d_i \in \{1, \ldots, d\}$ measurements. Here $d \in \mathbb{N}$ denotes the maximum number of measurements per individual observed. For two different individuals the jth event does not necessarily need to have occurred at the same time t_j. We denote by n_j the number of observations of length j, $j = 1, \ldots, d$, where n_j is zero if $\mathcal{Y}$ contains no observations of length j. We divide now the data set into subsets of groups of individuals with the same number of measurements. For $j = 1, \ldots, d$, we summarize the observations of group j as $\mathcal{Y}^j = \{\mathbf{y}^i \mid i \in I_j\}$, where the corresponding index set is defined as $I_j = \{i \mid \mathbf{y}^i \in \mathbb{R}^j\}$. Table 1 illustrates the above notation and data structure for an exemplary data set of size $n = 9$, where the maximum number of measurements per individual is $d = 4$ and we have $n_1 = 0$ individuals with 1 measurement, $n_2 = 3$ individuals with 2 measurements, $n_3 = 2$ individuals with 3 measurements and $n_4 = 4$ individuals with 4 measurements. Consequently, $I_1 = \emptyset$, $I_2 = \{1, 2, 3\}$, $I_3 = \{4, 5\}$ and $I_4 = \{6, 7, 8, 9\}$.

observations		measurements			
		1	2	3	4
$\mathcal{Y}^2 = \{\mathbf{y}^i \mid i \in I_2\}$	$\mathbf{y}^1$	*	*		
	$\mathbf{y}^2$	*	*		
	$\mathbf{y}^3$	*	*		
$\mathcal{Y}^3 = \{\mathbf{y}^i \mid i \in I_3\}$	$\mathbf{y}^4$	*	*	*	
	$\mathbf{y}^5$	*	*	*	
$\mathcal{Y}^4 = \{\mathbf{y}^i \mid i \in I_4\}$	$\mathbf{y}^6$	*	*	*	*
	$\mathbf{y}^7$	*	*	*	*
	$\mathbf{y}^8$	*	*	*	*
	$\mathbf{y}^9$	*	*	*	*

Table 1: Grouping of an exemplary data set of size $n = 9$ with $d = 4$, $n_2 = 3$, $n_3 = 2$ and $n_4 = 4$. Stars indicate observed events.

Having Sklar's Theorem (Sklar, 1959) in mind, we follow a two-stage approach, also referred to as the Inference Functions for Margins (IFM) method (cf. Joe, 1997, Section 10.1): First we use the probability integral transform and apply the univariate marginal distributions F_j^i to the measurements $y_j^i \in \mathbb{R}$ in order to transform them to measurements $u_j^i := F_j^i(y_j^i) \in [0, 1]$ to the uniform scale, $j = 1, \ldots, d_i$ and $i = 1, \ldots, n$. Then we model the dependence structure of the resulting uniform scale data utilizing a copula. In the following sections we will use a notation for the copula data that is similar to the one for the original data. The copula data $\mathcal{U} = \{\mathbf{u}^1, \ldots, \mathbf{u}^n\}$ consists of the observations $\mathbf{u}^i = (u_1^i, \ldots, u_{d_i}^i)^\top \in [0, 1]^{d_i}$, $i = 1, \ldots, n$. Again, we form groups

$\mathcal{U}^j = \{\mathbf{u}^i \mid i \in I_j\}$ containing all observations of length j, $j = 1, \ldots, d$. Since individuals with only one measurement do not contribute to the dependence structure we will only consider $\mathcal{U}^2, \ldots, \mathcal{U}^d$. Thus we can assume that $n_1 = 0$, i.e. $\mathcal{U}^1 = \emptyset$, without losing generality. Of course, in practice the distribution functions F_j^i are usually not known and need to be estimated (see Section 4).

2.2   D-vine based dependence model

D-vine copulas

Since we will use D-vine copulas for modeling the dependence we first give a short introduction to this model class. Vines have been introduced by Bedford and Cooke (2002). They are graphical models that can be used to construct a multivariate copula density as a product over bivariate building blocks, so-called pair-copulas. Since Aas et al. (2009) presented statistical inference methods for vine copulas, their popularity has increased drastically. D-vines are a subclass of vines representing a sequential structure. They are frequently used (e.g. Ren et al., 2014; Kim et al., 2013; Czado et al., 2011) because of their flexibility and interpretability.

If a continuous random vector $\mathbf{U}_{1:d} = (U_1, \ldots, U_d)^\top$ with uniform marginal distributions follows a D-vine copula density with order 1–2–$\ldots$–d, then, using the notation of Czado (2010), the density can be written as

$$c_{1:d}(u_1, \ldots, u_d) = \prod_{\ell=1}^{d-1} \prod_{k=1}^{d-\ell} c_{k,k+\ell;(k+1):(k+\ell-1)}\big(C_{k|(k+1):(k+\ell-1)}(u_k|u_{k+1}, \ldots, u_{k+\ell-1}),$$

$$C_{k+\ell|(k+1):(k+\ell-1)}(u_{k+\ell}|u_{k+1}, \ldots, u_{k+\ell-1}); u_{k+1}, \ldots, u_{k+\ell-1}\big). \tag{2.1}$$

Here $c_{k,k+\ell;(k+1):(k+\ell-1)}(\,\cdot\,,\cdot\,; u_{k+1}, \ldots, u_{k+\ell-1})$ is the bivariate copula density associated with the distribution of $(U_k, U_{k+\ell})^\top$ given $(U_{k+1}, \ldots, U_{k+\ell-1})^\top = (u_{k+1}, \ldots, u_{k+\ell-1})^\top$ and $C_{k|(k+1):(k+\ell-1)}(\,\cdot\,|u_{k+1}, \ldots, u_{k+\ell-1})$ is the distribution function of the conditional distribution of U_k given $(U_{k+1}, \ldots, U_{k+\ell-1})^\top = (u_{k+1}, \ldots, u_{k+\ell-1})^\top$, $\ell = 1, \ldots, d-1$ and $k = 1, \ldots, d-\ell$. The corresponding graphical interpretation is the tree representation, where the pair-copulas occurring in tree j have a conditioning set of size $j-1$, $j = 1, \ldots, d-1$. For $d = 4$ this concept is illustrated in Figure 1.

Many authors make the so-called simplifying assumption that the pair-copulas $c_{k,k+\ell;(k+1):(k+\ell-1)}(\,\cdot\,,\cdot\,; u_{k+1}, \ldots, u_{k+\ell-1})$ do not depend on values of the conditioning variables $u_{k+1}, \ldots, u_{k+\ell-1}$. More detailed investigations of this assumption can for example be found in Hobæk Haff et al. (2010), Acar et al. (2012), Stöber et al. (2013), Spanhel and Kurz (2015) and Killiches et al. (2017). We also make this assumption in order to ease inference later on although we could set up our model without it as well.

In the following we will assume a parametric model such that a D-vine copula can be identified by the set of pair-copula families $\mathcal{C} = (c_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d - \ell$ and $\ell = 1, \ldots, d-1)$ and the set of associated parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d-\ell$ and $\ell = 1, \ldots, d-1)$. In general, non-parametric pair-copulas could also be used (see Nagler and Czado, 2016).

A convenient property is that D-vine models are nested in the sense that the pair-copulas needed to describe the dependence of variables 1 to j are contained in the model describing the dependence of variables 1 to $j+1$, $j < d$. This is illustrated in Figure 1.

Model

Since the data has been obtained from repeated measurements there exists a clear sequential or temporal ordering. This immediately suggests the use of D-vine copulas with order 1–2–...–d (Smith et al., 2010; Smith, 2015; Nai Ruscone and Osmetti, 2017). Therefore, as a general approach, we assume parametric simplified D-vine models (cf. Equation (2.1)) for the copula densities of all groups $j = 2, \ldots, d$. Of course, we only consider groups for which we have observations. The copula density $c^j_{1:j}$ of group j then can be described with the help of the set of the $j(j-1)/2$ pair-copula families

$$\mathcal{C}^j = (c^j_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, j - \ell \text{ and } \ell = 1, \ldots, j - 1)$$

and the set of corresponding parameters

$$\boldsymbol{\theta}^j = (\boldsymbol{\theta}^j_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, j - \ell \text{ and } \ell = 1, \ldots, j - 1)$$

for $j = 2, \ldots, d$ with a non-empty $\mathcal{U}^j$. For the estimation of $\mathcal{C}^j$ and $\boldsymbol{\theta}^j$ we set up the likelihood, which is based on the subset of $\mathcal{U}$ containing the observations of length j. The resulting likelihood and log-likelihood can be written as

$$L_j(\mathcal{C}^j, \boldsymbol{\theta}^j \mid \mathcal{U}^j) = \prod_{i \in I_j} c^j_{1:j}(u^i_1, \ldots, u^i_j \mid \mathcal{C}^j, \boldsymbol{\theta}^j)$$

and

$$\log L_j(\mathcal{C}^j, \boldsymbol{\theta}^j \mid \mathcal{U}^j) = \sum_{i \in I_j} \log c^j_{1:j}(u^i_1, \ldots, u^i_j \mid \mathcal{C}^j, \boldsymbol{\theta}^j),$$

respectively. Consequently, the log-likelihood of the general model is given by

$$\log L(\mathcal{C}^2, \ldots, \mathcal{C}^d, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^d \mid \mathcal{U}) = \sum_{j=2}^{d} \log L_j(\mathcal{C}^j, \boldsymbol{\theta}^j \mid \mathcal{U}^j). \tag{2.2}$$

For future reference we call this *Model A*. It is obvious by construction that the models for different groups can be estimated independently from each other since there are no intersections between the groups, neither regarding data nor pair-copula families or parameters. From a practical point of view this would correspond to the assumption that the dependence structure of two groups can be completely different such that an individual for whom we have observed j events have nothing in common with those who have had $j + 1$ events. However, one can argue that an individual from group j is basically a member of group $j + 1$ for whom the $(j + 1)$st measurement has not been observed yet. The underlying random mechanism (i.e. the copula), however, should be the same or at least share some properties. Therefore, it makes sense to impose more restrictions on the set of pair-copula families and the associated parameters. For example, one could assume that all groups share the same pair-copula families and only the parameters can differ between the groups. The most sensible and interesting case—which we will pursue in the following—is the one that all groups have the same pair-copula families and parameters, i.e. for all $j = 2, \ldots, d$ we have

$$\begin{aligned}
c^j_{k,k+j;(k+1):(k+j-1)} &= c_{k,k+j;(k+1):(k+j-1)}, \\
\boldsymbol{\theta}^j_{k,k+j;(k+1):(k+j-1)} &= \boldsymbol{\theta}_{k,k+j;(k+1):(k+j-1)}
\end{aligned} \tag{2.3}$$

for $k = 1, \ldots, j - \ell$ and $\ell = 1, \ldots, j - 1$. We will refer to this model as *Model B*. Using the same families and parameters for all groups implies that the D-vine describing

the dependence pattern of group j is a sub-vine of the vine of groups $j + 1, \ldots, d$. In particular, the vine copula density of group j is simply the multivariate marginal density $c_{1:j}$ of the density $c_{1:d}$ of the largest group d. Consequently, $c_{1:d}$ describes the full model, from which the models of all smaller groups can be explicitly derived. Thus, the corresponding log-likelihood only depends on one set of $d(d-1)/2$ pair-copula families $\mathcal{C} = (c_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d - \ell \text{ and } \ell = 1, \ldots, d-1)$ and the set of corresponding parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d - \ell \text{ and } \ell = 1, \ldots, d-1)$.

Example

In order to illustrate the above concept we will now look at the example with at most $d = 4$ repeated measurements. Assume we have (up to) four-dimensional repeated measurement data $\mathcal{U} = \{\mathbf{u}^1, \ldots, \mathbf{u}^n\}$ of size $n = n_2 + n_3 + n_4$ ordered as described in Section 2.1, which can be partitioned into groups 2, 3 and 4 by defining $\mathcal{U}^j = \{u^i \mid i \in I_j\}$, $j = 2, 3, 4$, where $I_2 = \{i \mid \mathbf{u}^i \in \mathbb{R}^2\}$, $I_3 = \{i \mid \mathbf{u}^i \in \mathbb{R}^3\}$ and $I_4 = \{i \mid \mathbf{u}^i \in \mathbb{R}^4\}$. The model and hence the log-likelihood depends on the set of the six pair-copulas $\mathcal{C} = (c_{1,2}, c_{2,3}, c_{3,4}, c_{1,3;2}, c_{2,4;3}, c_{1,4;2,3})$ and the associated parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,3}, \boldsymbol{\theta}_{3,4}, \boldsymbol{\theta}_{1,3;2}, \boldsymbol{\theta}_{2,4;3}, \boldsymbol{\theta}_{1,4;2,3})$. Figure 1 shows a schematic representation of the full model $c_{1:4}$ with its pair-copulas and parameters. The nodes represent the measurements. Above and below each edge the associated pair-copula and the observations that can be used for estimation are denoted, respectively. The sub-vines for $c_{1:2}$ and $c_{1:3}$ are highlighted by different color intensities of the nodes and line types of the edges. The resulting log-likelihood is given by

$$
\begin{aligned}
\log L(\mathcal{C}, \boldsymbol{\theta} \mid \mathcal{U}) = &\sum_{i \in I_2} \log c_{1:2}(u_1^i, u_2^i \mid c_{1,2}, \boldsymbol{\theta}_{1,2}) \\
&+ \sum_{i \in I_3} \log c_{1:3}(u_1^i, u_2^i, u_3^i \mid c_{1,2}, c_{2,3}, c_{1,3;2}, \boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,3}, \boldsymbol{\theta}_{1,3;2}) \\
&+ \sum_{i \in I_4} \log c_{1:4}(u_1^i, u_2^i, u_3^i, u_4^i \mid c_{1,2}, c_{2,3}, c_{3,4}, c_{1,3;2}, c_{2,4;,3}, c_{1,4;2,3}, \\
&\qquad\qquad\qquad \boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,3}, \boldsymbol{\theta}_{3,4}, \boldsymbol{\theta}_{1,3;2}, \boldsymbol{\theta}_{2,4;,3}, \boldsymbol{\theta}_{1,4;2,3})
\end{aligned}
\tag{2.4}
$$

Using the vine decomposition from Equation (2.1) for $c_{1:2}$, $c_{1:3}$ and $c_{1:4}$, the log-likelihood associated with data $\mathcal{U}$ (Equation (2.4)) can be re-written as

$$
\begin{aligned}
\log L(\mathcal{C}, \boldsymbol{\theta} \mid \mathcal{U}) = \\
&\sum_{i \in I_2 \cup I_3 \cup I_4} \log c_{1,2}(u_1^i, u_2^i; \boldsymbol{\theta}_{1,2}) \\
&+ \sum_{i \in I_3 \cup I_4} \left[\log c_{2,3}(u_2^i, u_3^i; \boldsymbol{\theta}_{2,3}) + \log c_{1,3;2}(C_{1|2}(u_1^i|u_2^i; \boldsymbol{\theta}_{12}), C_{3|2}(u_3^i|u_2^i; \boldsymbol{\theta}_{23}); \boldsymbol{\theta}_{1,3;2}) \right] \\
&+ \sum_{i \in I_4} \left[\log c_{3,4}(u_3^i, u_4^i; \boldsymbol{\theta}_{3,4}) + \log c_{2,4;3}(C_{2|3}(u_2^i|u_3^i; \boldsymbol{\theta}_{2,3}), C_{4|3}(u_4^i|u_3^i; \boldsymbol{\theta}_{4,3}); \boldsymbol{\theta}_{2,4;3}) \right. \\
&\qquad + \log c_{1,4;2,3}(C_{1|3;2}(C_{1|2}(u_1^i|u_2^i; \boldsymbol{\theta}_{1,2}) \mid C_{3|2}(u_3^i|u_2^i; \boldsymbol{\theta}_{2,3}); \boldsymbol{\theta}_{1,3;2}), \\
&\qquad\qquad \left. C_{4|2;3}(C_{2|3}(u_2^i|u_3^i; \boldsymbol{\theta}_{2,3}) \mid C_{4|3}(u_4^i|u_3^i; \boldsymbol{\theta}_{4,3}); \boldsymbol{\theta}_{2,4;3}); \boldsymbol{\theta}_{1,4;2,3}) \right]
\end{aligned}
\tag{2.5}
$$
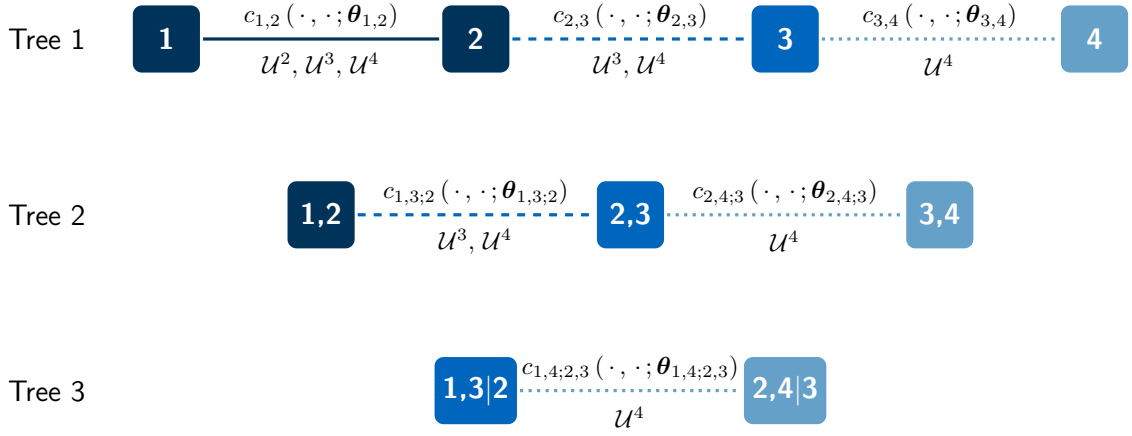
Figure 1: Illustration of the four-dimensional D-vine describing the components of the dependence structure of the full model $c_{1:4}$ (dark, medium and light). The sub-vines for $c_{1:2}$ (dark) and $c_{1:3}$ (dark and medium) are highlighted by different color intensities of the nodes and line types of the edges. Above and below each edge the associated pair-copula and the observations that can be used for estimation are denoted, respectively.

For the general case of Model A (Equation (2.2)) we saw that the pair-copulas and parameters corresponding to group j can be estimated independently from those of the remaining groups and only depend on the data contained in $\mathcal{U}^j$. Looking at Equation (2.5) (corresponding to Model B) it immediately becomes clear that assuming the pair-copulas and parameters are the same for all groups has changed this phenomenon. The D-vines describing the densities $c_{1:2}$ and $c_{1:3}$ are nested sub-vines of the full model $c_{1:4}$, which can easily be understood from Figure 1: The dark nodes (and solid edges) correspond to $c_{1:2}$; adding the medium colored nodes (and dashed edges) results in the model of $c_{1:3}$; incorporating also the light nodes (and dotted edges) yields the full model for $c_{1:4}$. Therefore, when it comes to estimation we see for example that not only the observations belonging to $\mathcal{U}^2$ but also those from $\mathcal{U}^3$ and $\mathcal{U}^4$ (i.e. the entire sample $\mathcal{U}$) have an influence on the estimate $c_{1,2}$ and $\boldsymbol{\theta}_{1,2}$. Thus this increases the accuracy of the estimation compared to the approach from Model A.

The assumption of common pair-copula families and parameters for all groups come with the advantages of better interpretability, less parameters and higher estimation accuracy.

Missing values

In practice, unfortunately, data do not always look exactly the way we described it in Section 2.1. Sometimes there are missing values in the data. For example, there might be individuals for whom the first, third, fourth and fifth measurement are available but the second one is missing. Such situation can occur for various reasons, e.g. a patient skips a measurement date due to illness, measuring instruments have problems causing a loss of the result or data is simply not reported due to human failure. Moreover, there might be (non-informative) dropouts, i.e. individuals with measurements only up to a certain time, e.g. caused by relocation of a patient to another city. For many model classes such observations cannot be used at all and have to be removed from the data set for model estimation. This way the sample size is decreased and information is lost. For Model B, however, observations with missing values can still be used (assuming they are missing at random). The information gained from our exemplary individual

with measurements 1, 3, 4, 5 still contributes to the estimation of $c_{3,4}$, $c_{4,5}$ and $c_{3,5;4}$ (and of course to the estimation of the marginal distributions 1, 3, 4, 5). Since the missing second measurement is needed for the estimation of the remaining pair-copulas, this individual cannot be used in order to estimate them. Nevertheless, we prevent the loss of the individual's entire information. In order to include observations with missing value into our model we simply have to modify the log-likelihood such that the sums of the log-likelihood of each pair-copula includes all observations for whom the necessary measurements are available. For the sake of notation we will stick to the formulation of Model B as above, keeping in mind that missing values can also be handled by our approach.

Conditional prediction

Further, we can use our repeated measurement data model for prediction. In many applications it can be interesting to have a prediction for the size of an upcoming measurement. For instance, having proper estimates for future claims can be a competitive advantage for the risk management department of an insurance company.

For a d-dimensional model, consider an individual i that has had $d_i < d$ measurements so far, i.e. $\mathbf{y}^i = (y_1^i, \ldots, y_{d_i}^i)^\top$. We are now interested in the distribution of the next measurement $d_i + 1$. Since $d_i + 1 \leq d$, the sub-vine describing the dependence of events 1 to $d_i + 1$ can be extracted from the full model. We consider the conditional distribution function $F_{d_i+1|1:d_i}^i(\cdot|y_1^i, \ldots, y_{d_i}^i)$. Joe (1997) was the first to show that there exists a recursive representation for such conditional distribution functions. This way one obtains a closed-form expression of the conditional distribution function solely based on the pair-copulas specified in the D-vine (and the univariate marginals, of course) if the variable to be predicted is a leaf in the first tree. In our case, $d_i + 1$ is in fact a leaf in the first tree of the D-vine on nodes 1 to $d_i + 1$. Thus, we know $F_{d_i+1|1:d_i}^i(\cdot|y_1^i, \ldots, y_{d_i}^i)$ analytically and can further simulate from it. For example, we can express $F_{4|1,2,3}^i$ in the following way:

$$F_{4|1,2,3}^i(y_4^i|y_1^i, y_2^i, y_3^i) = C_{4|1;23}\left(C_{4|2;3}\left(C_{4|3}(F_4^i(y_4^i)|F_3^i(y_3^i))\big|C_{2|3}(F_2^i(y_2^i)|F_3^i(y_3^i))\right)\big|\right.$$
$$\left. C_{1|3;2}\left(C_{1|2}(F_1^i(y_1^i)|F_2^i(y_2^i))\big|C_{3|2}(F_3^i(y_3^i)|F_2^i(y_2^i))\right)\right).$$

Further, the conditional quantile function can be expressed in general as

$$\begin{aligned} q_\alpha(y_1^i, \ldots, y_{d_i}^i) &= (F_{d_i+1|1:d_i}^i)^{-1}(\alpha|y_1^i, \ldots, y_{d_i}^i) \\ &= (F_{d_i+1}^i)^{-1}\left(C_{d_i+1|1:d_i}^{-1}(\alpha|F_1^i(y_1^i), \ldots, F_{d_i}^i(y_{d_i}^i))\right) \end{aligned} \tag{2.6}$$

and is of great interest in order to determine upper and lower bounds of a confidence interval. Kraus and Czado (2017a) show that inversion also yields a closed-form expression for the conditional quantile function solely based on the specified pair-copulas and marginals. Thus, we can determine arbitrary conditional quantiles for the size of measurement $d_i + 1$. For example, for financial applications it might be interesting to obtain a conditional 99%-value at risk, i.e. the conditional 99%-quantile, for the size of individual i's next measurement.

In order to be able to perform statistical inference of any kind with our D-vine model we first have to estimate the pair-copula families and associated parameters. Section 4 will present two estimation approaches. First, however, we will introduce linear mixed model and illustrate how they are connected to our proposed D-vine based model in Section 3.

3 Connection between the D-vine based model and linear mixed models

Probably the most popular models for longitudinal data are linear mixed models. In this section we will give a short introduction to this model class and show how they are connected to our approach from Section 2.

3.1 Linear mixed models for repeated measurements

Linear mixed models have been discussed in detail by many authors, e.g. in Diggle (2002), Verbeke and Molenberghs (2009) and Fahrmeir et al. (2013). Describing the outcome of repeated measurements j, $j = 1, \ldots, d_i$, for individuals i, $i = 1, \ldots, n$ as responses Y_j^i, they extend linear models by including random effects $\boldsymbol{\gamma}_i \in \mathbb{R}^q$ to the fixed (i.e. non-random) effects $\beta \in \mathbb{R}^p$, $p, q \in \mathbb{N}$. These random effects, unlike the fixed effects, are different for each individual. The covariate vectors $\mathbf{x}_{i,j} \in \mathbb{R}^p$ and $\mathbf{z}_{i,j} \in \mathbb{R}^q$ are associated to the fixed and random effects, respectively.

For $i = 1, \ldots, n$ and $j = 1, \ldots, d_i$, the jth measurement for individual i is assumed to decompose to

$$Y_j^i = \mathbf{x}_{i,j}^\top \boldsymbol{\beta} + \mathbf{z}_{i,j}^\top \boldsymbol{\gamma}_i + \varepsilon_{i,j}, \tag{3.7}$$

where the vector of random effects $\boldsymbol{\gamma}_i \sim \mathcal{N}_q(\mathbf{0}, D)$ is normally distributed with zero expectation covariance matrix $D \in \mathbb{R}^{q \times q}$ and the error vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,d_i})^\top \sim \mathcal{N}_{d_i}(\mathbf{0}, \Sigma_i)$ also follows a centered normal distribution with covariance matrix $\Sigma_i \in \mathbb{R}^{d_i \times d_i}$. Further, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are assumed to be independent. Hence,

$$Y_j^i \sim \mathcal{N}(\mathbf{x}_{i,j}^\top \boldsymbol{\beta}, \phi_{i,j}^2) \tag{3.8}$$

with standard deviation $\phi_{i,j} := \left(\mathbf{z}_{i,j}^\top D \mathbf{z}_{i,j} + \sigma_{i,j}^2\right)^{1/2}$, where $\sigma_{i,j}^2 := \mathrm{Var}(\varepsilon_{i,j})$. Using the notation

$$X_i := \begin{pmatrix} \mathbf{x}_{i,1}^\top \\ \vdots \\ \mathbf{x}_{i,d_i}^\top \end{pmatrix} \in \mathbb{R}^{d_i \times p}, \quad Z_i := \begin{pmatrix} \mathbf{z}_{i,1}^\top \\ \vdots \\ \mathbf{z}_{i,d_i}^\top \end{pmatrix} \in \mathbb{R}^{d_i \times q}, \quad \mathbf{Y}^i := \begin{pmatrix} Y_1^i \\ \vdots \\ Y_{d_i}^i \end{pmatrix} \in \mathbb{R}^{d_i}$$

we can represent the vector of all measurements belonging to individual i as follows:

$$\mathbf{Y}^i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i. \tag{3.9}$$

We see that due to the independence assumptions of $\boldsymbol{\gamma}_i$ and $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$, there exists a correlation between measurements of one individual but measurements of different individuals are independent. Further, the joint distribution of $\mathbf{Y}^i$ can be determined to be

$$\mathbf{Y}^i \sim \mathcal{N}_{d_i}(X_i \boldsymbol{\beta}, Z_i D Z_i^\top + \Sigma_i) \tag{3.10}$$

and $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are independent. The fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{\gamma}_i$ as well as the parameters of the covariance matrices D and Σ_i, $i = 1, \ldots, n$, can be estimated using maximum likelihood estimation as described for example in Diggle (2002) and Fahrmeir et al. (2013).

Linear mixed models are very popular in practice since they are easy to handle and interpret. Further, observations with missing data can also be used for ML estimation as long as the values are missing at random (see e.g. McCulloch et al., 2011; Ibrahim and Molenberghs, 2009).

3.2 Aligning linear mixed models and the D-vine based approach

Equation (3.10) implies that all univariate marginal distributions are normal distributions. Further, the dependence structure is Gaussian and can vary from individual to individual since the correlation matrix R_i of $\mathbf{Y}^i$ is given by

$$R_i := \operatorname{Cor}(\mathbf{Y}^i) = \operatorname{diag}(\phi_{i,1}^{-1}, \ldots, \phi_{i,d_i}^{-1}) \left(Z_i D Z_i^\top + \Sigma_i \right) \operatorname{diag}(\phi_{i,1}^{-1}, \ldots, \phi_{i,d_i}^{-1}),$$

where $\phi_{i,j}$ is the standard deviation of Y_j^i, $j = 1, \ldots, d_i$, $i = 1, \ldots, n$. In practice, however, this would make estimation infeasible since the number of parameters would be too large; in many cases one would even have more parameters than observations. Therefore, structural assumptions are made, especially for $\Sigma_i \in \mathbb{R}^{d_i \times d_i}$, in order reduce the number of parameters to be estimated.

In Section 2.2 we assumed that the dependence structure is basically the same for all individuals and only differs due to the number of measurements d_i that individual i has had so far. In order to obtain the same for linear mixed models, we simply have to require the following homogeneity condition:

Homogeneity condition: We call correlation matrices R_i *homogeneous* if they are the same for all individuals $i = 1, \ldots, n$ except for the dimension, i.e. $R_i = (r_{k,\ell})_{k,\ell=1}^{d_i} \in \mathbb{R}^{d_i \times d_i}$ is a $(d_i \times d_i)$-submatrix of a correlation matrix $R = R_d = (r_{k,\ell})_{k,\ell=1}^{d} \in \mathbb{R}^{d \times d}$.

This condition is in particular fulfilled if the covariance matrices of the errors $\Sigma_i \in \mathbb{R}^{d_i \times d_i}$ and the design matrices of the random effects $Z_i \in \mathbb{R}^{d_i \times q}$ are constant in i except for the dimension. Despite being a restriction, linear mixed models meeting this requirement still comprise a wide range of models used in practice. The assumption on the covariance matrices Σ_i is for example fulfilled if errors

- are assumed to be i.i.d., i.e. the (k, ℓ)th entry of Σ_i is given by $\sigma^2 \mathbf{1}\{k = \ell\}$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function;

- exhibit a compound symmetry structure, i.e. the (k, ℓ)th entry of Σ_i is $\sigma^2 \rho^{\mathbf{1}\{k \neq \ell\}}$ for some $\rho \in (-1, 1)$;

- follow an autoregressive structure of order 1 (AR(1)), i.e. the (k, ℓ)th entry of Σ_i is given by $\sigma^2 \rho^{|k-\ell|}$ for some $\rho \in (-1, 1)$;

- have an exponential decay structure, i.e. the (k, ℓ)th entry of Σ_i is given by $\sigma^2 \exp\{-|k - \ell|/r\}$, where $r > 0$ is the constant "range" parameter.

These are typical simplifications that are made anyway for modeling longitudinal data in most applications if the number of individuals is large with respect to the number of measurements. The assumption on the design matrices Z_i is also often satisfied, e.g. for the popular class of so-called *random intercept models*, where $Z_i = (1, \ldots, 1)^\top \in \mathbb{R}^{d_i \times 1}$ for $j = 1, \ldots, d_i$ and $i = 1, \ldots, n$. Further, the assumption includes any model where the covariates associated with the random effect only depend on the (common) measurement times t_j, $j = 1, \ldots, d$, i.e. for example $Z_i = (t_1, \ldots, t_{d_i})^\top \in \mathbb{R}^{d_i \times 1}$ or more generally $Z_i = (h(t_1), \ldots, h(t_{d_i}))^\top \in \mathbb{R}^{d_i \times 1}$ for some function $h \colon \mathbb{R} \to \mathbb{R}$. Thus, assuming that Z_i only depends on the number of measurements d_i for individual i is also not uncommon such that there is in fact a wide class of linear mixed models sharing the property that the correlation matrix R_i of $\mathbf{Y}^i$ only depends on the number of measurements.

If R_i is homogeneous in i, we have that all individuals i share the same Gaussian dependence structure, i.e. correlation matrix. This scenario is a special case of the D-vine based model since we can represent any Gaussian correlation matrix using a D-vine with Gaussian pair-copulas and the corresponding (partial) correlations as parameters (see for example Stöber et al., 2013, Theorem 4.1). The univariate margins F_j^i can be chosen arbitrarily for the copula approach such that we can simply use $\mathcal{N}(\mathbf{x}_{i,j}^\top \boldsymbol{\beta}, \phi_{i,j}^2)$-margins (cf. Equation (3.8)) to end up with a model describing the same joint distribution of $\mathbf{Y}^i$ as the corresponding linear mixed model (Equation (3.10)). Since we can use arbitrary distributions for the margins and/or any D-vine copula for the dependence structure, our approach can be seen as an extension of linear mixed models with common correlation structure for all individuals. Figure 2 illustrates the link between our D-vine based model and linear mixed models.
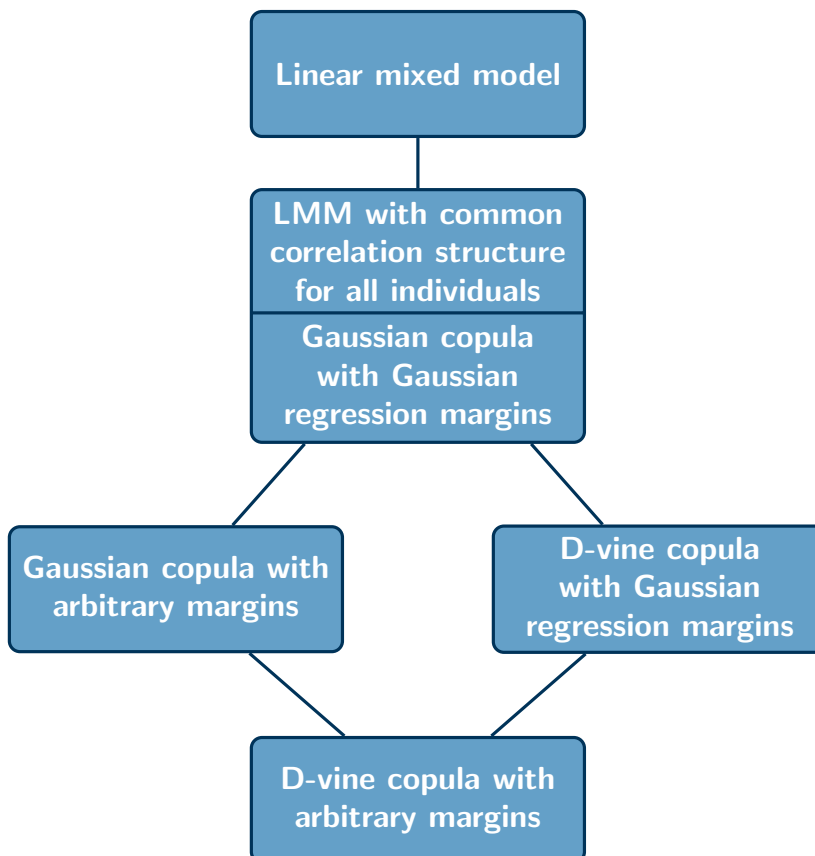
Figure 2: Flow chart illustrating how the D-vine based model is linked to linear mixed models.

For the application in Section 6 we will compare how well both model classes perform fitting real life data.

4   Estimation methods for the D-vine based model

4.1   Marginal modeling

Although our focus is on dependence modeling, we will briefly discuss how the univariate marginal models for Y_j^i, i.e. F_j^i, can be estimated. In general the choice of marginal models is completely arbitrary. They can be parametric or non-parametric. The most common situation for repeated measurements is that in addition to the measurement

data $\mathcal{Y}$ itself further covariates are known for each individual and measurement. Therefore, regression models such as linear (LMs) or generalized linear models (GLMs) can be applied. In this case, $F_j^i(\cdot) = F_j(\cdot \,|\, \mathbf{x}_{i,j})$ depends on the individual's associated covariates $\mathbf{x}_{i,j} \in \mathbb{R}^p$, where $p \in \mathbb{N}$ is the number of covariates used in the model. In our application (Section 6) we will fit linear models to the margins. In order to avoid overfitting we determine which covariates to include into the models by BIC-based forward selection, i.e. covariates are added to the model until BIC (Bayesian information criterion, see Schwarz, 1978) cannot be improved anymore. Our focus, however, is rather to develop a flexible model describing the dependence structure that is present in the data $\mathcal{Y}$ such that we will not further elaborate on how to estimate the univariate marginal distributions.

4.2 Dependence modeling

Assume we have estimated the marginal distributions and obtained (pseudo-)copula data by applying the estimated distribution functions $\hat{F}_j^i$ to the measurements, i.e. $\hat{u}_j^i := \hat{F}_j^i(y_j^i)$. We now use the transformed data as a copula sample to estimate the underlying dependence structure. Section 2 has shown that D-vine copulas are suited for modeling the dependence structure being present in repeated measurement data. Model B (Equation (2.3)) was preferable since it is easier to interpret and estimate. Further predictions for not yet observed measurements can be made. The aim of the methods presented in Section 4 is to find estimates for the set of pair-copula families $\mathcal{C} = (c_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d-\ell \text{ and } \ell = 1, \ldots, d-1)$ and the set of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{k,k+\ell;(k+1):(k+\ell-1)} \mid k = 1, \ldots, d-\ell \text{ and } \ell = 1, \ldots, d-1)$ corresponding to Model B from Section 2, where d is the maximal number of observed events per observation. We will present two approaches: a standard joint maximum-likelihood estimator and a sequential method. Since we want to choose both parameters and families for each pair-copula we will select from a set of m bivariate candidate family types $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$, where each member $\gamma \in \Gamma$ has its own space of admissible parameters $\Omega(\gamma)$.

Joint maximum-likelihood approach
The canonical approach in order to find optimal pair-copula families and parameters would be to use maximum-likelihood estimation. In Section 2 we have already determined the log-likelihood. Since the families specify which parameters are admissible, finding the optimal families and parameters is divided into two steps: For each combination of families we have to determine the maximum-likelihood estimate of parameters; then we select the one combination with the overall highest likelihood. This way we find the best D-vine model with regards to likelihood optimization. However, since there are $|\Gamma| = m$ candidates for each of the $d(d-1)/2$ pair-copula families, we have to perform $m^{d(d-1)/2}$ times an at least $(d(d-1)/2)$-dimensional optimization (some families like the t-copula may have more than one parameter). It is obvious that this can very quickly become computationally infeasible if the number of candidate families m is high and—especially—if the dimension d gets large.

Of course, the possibly large number of parameters to be estimated is a general problem in the statistical analysis of vine copulas. Therefore, Aas et al. (2009) (for D-vines) and later Dißmann et al. (2013) (for general vine copulas) developed a sequential tree-by-tree selection algorithm facilitating vine copula model estimation up to very high dimensions. Dißmann's algorithm is commonly used to fit the vine's model struc-

ture, pair-copula families and parameters but it can also only be used for the selection of families and parameters only if we have a fixed tree structure (e.g. a D-vine). The difference to the classical situation which we face when we want to estimate a vine copula is that our observations have different lengths.

Sequential approach

Inspired by Dißmann's algorithm we want to fit the pair-copula families and the associated parameters of the D-vine to a repeated measurement data set using a sequential approach. Given classical data, Dißmann's algorithm starts with the estimation of the first tree and estimates the unconditional pair-copulas (and their parameters) via maximum-likelihood estimation. Then the observations are transformed into pseudo-observations needed for the estimation of the second tree using the estimated pair-copulas of tree 1. Continuing this way the vine is built up tree-by-tree.

In the presence of repeated measurement data, however, we can pursue a very similar strategy. The only difference is that we estimate each pair-copula (and its parameter(s)) only based on the available full observation. All pair-copulas to be estimated are of the form $c_{k,\ell;(k+1):(\ell-1)}$ with parameter $\boldsymbol{\theta}_{k,\ell;(k+1):(\ell-1)}$. When all observations are of the form $(u_1, \ldots, u_j)^\top$, i.e. there are no "gaps" between two observed events, we can use the information of observations with a minimum length of ℓ, i.e. all observations in $\bigcup_{j=\ell}^d \mathcal{U}^j$, for the estimation of $c_{k,\ell;(k+1):(\ell-1)}$ and $\boldsymbol{\theta}_{k,\ell;(k+1):(\ell-1)}$. Thus, we can maintain the basic scheme known from Dißmann's algorithm. With a slight modification of the data we are even able to use the function `RVineCopSelect` from the R library `VineCopula` (Schepsmeier et al., 2017) for our purpose, making our approach also very appealing from a practitioner's point of view.

Of course, this sequential approach can also be applied for data with missing values (Section 2.2, page 7). Then, for the estimation of each pair-copula $c_{k,\ell;(k+1):(\ell-1)}$ with associated parameter $\boldsymbol{\theta}_{k,\ell;(k+1):(\ell-1)}$ is performed using all observations for whom the necessary measurements $u_k, u_{k+1}, \ldots, u_l$ are available. The function `RVineCopSelect` can still be used in the presence of missing values.

The biggest advantage of being able to use sequential estimation approach is that we can estimate models at reasonable computational costs, even in high dimensions. Of course, the approach also works when using non-parametric pair-copulas or even non-simplified vine copulas. For details for estimating non-parametric and non-simplified vines we refer the reader to Nagler and Czado (2016) and Vatter and Nagler (2016), respectively. Yet, as already mentioned at the beginning, we focus on parametric simplified vine copulas here.

4.3   Model selection

In model selection one often wants to compare different fitted models. For this purpose the log-likelihood and log-likelihood based measures such as AIC (Akaike, 1998) and BIC (Schwarz, 1978), which penalize large numbers of parameters, are frequently applied. Whereas the penalty of the AIC only depends on the number of parameters in the model, that of BIC also depends on the sample size. In our case, however, it is not completely obvious what sample size to use. Therefore, we derive how the BIC for the D-vine based model including margins can be calculated in our situation. Proposition 4.1 shows that each parameter is to be weighted with the (logarithm of) the number of observations that directly contribute to its estimation. A proof can be found in Appendix A.

Proposition 4.1. Let $p_j \in \mathbb{N}$ be the number of parameters of the D-vine based model including margins restricted to the measurements 1 to j, $j = 1, \ldots, d$, and define $\Delta p_j := p_j - p_{j-1}$ for $j = 2, \ldots, d$ and $\Delta p_1 := p_1$. Further denote by $N_j = \sum_{k=j}^{d} n_k$ the number of individuals with at least j measurements. The BIC of the D-vine based model including margins is given by

$$\text{BIC} = -2 \log L(\hat{\boldsymbol{\theta}} \,|\, \mathcal{Y}) + \sum_{j=1}^{d} \Delta p_j \log(N_j).$$

Here, $\log L(\hat{\boldsymbol{\theta}} \,|\, \mathcal{Y}) = \log L(\hat{\boldsymbol{\theta}}_M \,|\, \mathcal{Y}) + \log L(\hat{\boldsymbol{\theta}}_C \,|\, \mathcal{U})$ is the log-likelihood of the fitted model including margins, i.e. the sum of the log-likelihood of the margins $\log L(\hat{\boldsymbol{\theta}}_M \,|\, \mathcal{Y})$ and the one of the copula $\log L(\hat{\boldsymbol{\theta}}_C \,|\, \mathcal{U})$ (which is the one of Model B from Section 2.2). Further, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_C)$ is the maximum-likelihood estimate for the set of all model parameters (associated with both the margins $\hat{\boldsymbol{\theta}}_M$ and the D-vine copula $\hat{\boldsymbol{\theta}}_C$).

Remark 4.2. Although this BIC adjustment was developed for the D-vine based model, it can also be used for certain types of linear mixed models due to the connection described in Section 3. For LMMs fulfilling the homogeneity condition the BIC can be determined with the formula from Proposition 4.1 if only individuals with j or more measurements contribute to the estimation of the parameters of the sub-model restricted to the first j measurements which were not already contained in the sub-model restricted to the first $j - 1$ measurements. This is for example the case if on the one hand no structural assumptions (besides homogeneity) are imposed on the covariance matrices of the random effects and the errors D and Σ_i and on the other hand the design matrices X_i have a form that allows for different marginal regression models for different measurements. For guaranteeing the latter each covariate is only allowed to be incorporated in one of the marginal regressions, i.e. the values of this covariate are zero for all other measurements; if a covariate still is to be included in more than one model, one simply splits up the covariate into several measurements-specific covariates that are non-zero only for one particular measurement. This way an own coefficient for one covariate can be estimated for different marginal models (if necessary).

5 Simulation study

In order to check that the sequential estimation approach from Section 4 works reasonably well, we perform a simulation study that is inspired by the data analyzed in Section 6.

Simulation setting

For a maximum number of measurements $d \in \{5, 10\}$, we generate d-dimensional data sets and prune them randomly to obtain an unbalanced setting. In this context pruning means that for each d-dimensional observation i we independently draw d_i from a discrete distribution on $\{2, \ldots, d\}$ and restrict this observation to its first d_i components. This way we mimic the nature of unbalanced repeated measurement data. In order to assess the implications of having only incomplete data we sequentially fit a D-vine copula to both the full and the pruned data set and compare the estimates.

To obtain data sets we consider randomly generate D-vine copulas with structure 1–2–$\ldots$–d. For this purpose, we rely on the method proposed in Joe (2006) to sample

14

Gaussian correlation matrices that are uniformly distributed over the space of valid correlation matrices. Conveniently, this method is already based on a vine decomposition: For each tree i, $i = 1, \ldots, d-1$, we generate the corresponding $d-i$ parameters associated to the Gaussian pair-copulas by drawing from a Beta$((d-i+1)/2, (d-i+1)/2)$ distribution and transforming the outcome linearly to $[-1, 1]$, resulting in a mean and mode of 0 and a variance of $1/(d-i+2)$. However, since we do not only want to consider Gaussian D-vines, we transform the Gaussian parameters to Kendall's τ values using the relationship $\tau = \frac{2}{\pi} \arcsin(\rho)$. Then, we randomly draw a pair-copula family for each pair-copula to be specified[1] and convert the Kendall's τ values to parameters of the respective families. For one-parametric families τ can directly be transformed to the parameter space. For two-parametric families there are infinitely many combinations of parameters resulting in the same Kendall's τ value. Therefore, we adopt the approach used in Kraus and Czado (2017b): draw the second parameter randomly[2] and determine the first parameter implicitly such that the two parameters imply the required Kendall's τ.

With the above procedure we generate $R = 1000$ D-vine copulas and simulate data sets of size $n \in \{200, 2000\}$. Then for each observation i we randomly determine its length $d_i \in \{2, \ldots, d\}$. For $d = 5$, the underlying distribution mimics the observed measurement rates of the data considered in Section 6. The exact proportions of individuals with a least j measurements would have been $100.0\%, 78.5\%, 58.5\%, 43.9\%$ for $j = 2, 3, 4, 5$, respectively. For $d = 10$, we extended the scenario of $d = 5$ accordingly. The distributions are given in Table 2.

j	2	3	4	5					
probability of $d_i = j$	20%	20%	15%	45%					
probability of $d_i \geq j$	100%	80%	60%	45%					

j	2	3	4	5	6	7	8	9	10
probability of $d_i = j$	10%	10%	10%	10%	10%	5%	5%	5%	35%
probability of $d_i \geq j$	100%	90%	80%	70%	60%	50%	45%	40%	35%

Table 2: Probability mass function and proportions of individuals with at least j measurements for the "pruning distribution" (top table: $d = 5$; bottom table: $d = 10$).

For both the full and the pruned data set we use the sequential algorithm implemented in `RVineCopSelect` (from the `VineCopula` library) to fit D-vine copulas. In order to assess how badly the loss of information affects the estimation we compare the resulting D-vines by considering each pair-copula separately. For this purpose, we consider the mean absolute difference between the Kendall's τ values $(\Delta_\tau := \frac{1}{R} \sum_{r=1}^{R} |\hat{\tau}_{\text{pruned}}(r) - \hat{\tau}_{\text{full}}(r)|)$, the lower and the upper tail dependence coefficients $(\Delta_{\lambda^s} := \frac{1}{R} \sum_{r=1}^{R} |\hat{\lambda}_{\text{pruned}}^s(r) - \hat{\lambda}_{\text{full}}^s(r)|$ for $s \in \{\ell, u\})$ of the two models. Comparing general strength of dependence and tail behavior enables us to assess how similar the fitted pair-copulas are.[3]

---

[1]The families are drawn uniformly from the ones available in the library `VineCopula`: Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8 and Tawn as well their rotations (see Schepsmeier et al., 2017, for details).

[2]The specific sampling distributions can be found in Appendix B of Kraus and Czado (2017b).

[3]Considering the percentage of cases where the same copula family is fitted would not be sensible since the number of candidate families is large and many of them, e.g. a Clayton and a survival Joe copula, can hardly be distinguished.

Results for $d = 5$

For $d = 5$, the absolute differences of Kendall's τ, lower and upper tail dependence coefficient (averaged over the $R = 1000$ data sets) are displayed for each of the 10 pair-copulas in Table 3, where the sample sizes are $n = 200$ and $n = 2000$, respectively. For $n = 200$, the 10 average absolute estimated Kendall's τ values for the full data set $(\frac{1}{R}\sum_{r=1}^{R}|\hat{\tau}_{\text{full}}(r)|)$ lie between 0.345 and 0.394; the 10 average estimated upper and lower tail dependence coefficients for the 10 pair-copulas are between 0.075 and 0.108 $(\frac{1}{R}\sum_{r=1}^{R}\hat{\lambda}_{\text{full}}^{\ell}(r))$ and 0.080 and 0.107 $(\frac{1}{R}\sum_{r=1}^{R}\hat{\lambda}_{\text{full}}^{u}(r))$, respectively. For $n = 2000$, the three ranges are fairly similar: $[0.338, 0.422]$, $[0.081, 0.992]$ and $[0.083, 0.107]$, respectively.

		$c_{1,2}$	$c_{2,3}$	$c_{3,4}$	$c_{4,5}$	$c_{1,3;2}$	$c_{2,4;3}$	$c_{3,5;4}$	$c_{1,4;2,3}$	$c_{2,5;3,4}$	$c_{1,5;2,3,4}$
$n = 200$	Δ_{τ}	0.000	0.016	0.026	0.035	0.017	0.027	0.036	0.032	0.043	0.058
	$\Delta_{\lambda^{\ell}}$	0.000	0.018	0.035	0.050	0.024	0.039	0.053	0.044	0.066	0.065
	$\Delta_{\lambda^{u}}$	0.000	0.022	0.030	0.041	0.024	0.037	0.058	0.048	0.067	0.065
$n = 2000$	Δ_{τ}	0.000	0.005	0.007	0.010	0.005	0.008	0.010	0.008	0.011	0.015
	$\Delta_{\lambda^{\ell}}$	0.000	0.004	0.008	0.012	0.009	0.011	0.016	0.008	0.021	0.015
	$\Delta_{\lambda^{u}}$	0.000	0.004	0.010	0.011	0.006	0.008	0.013	0.011	0.021	0.024

Table 3: Absolute differences of Kendall's τ, lower and upper tail dependence coefficient for each of the 10 pair-copulas, averaged over the $R = 1000$ data sets of size $n = 200$ and $n = 2000$, respectively.

We can see that even for a sample size of only $n = 200$ (see upper part of Table 3) the differences between the two estimates are relatively small. The largest absolute deviations are 0.058, 0.066 and 0.067 for τ, λ^{ℓ} and λ^{u}, respectively. The average absolute deviations 0.029 (τ), 0.039 (λ^{ℓ}) and 0.039 (λ^{u}), respectively. Of course, $c_{1,2}$ is always estimated equally in both cases since all pruned observations have minimum length of 2. We can observe what one would expect given that the number of observations with at least j measurements descends in j: Pair-copulas for whose estimation later measurements are needed exhibit larger deviations.

The results in the lower part of Table 3 (corresponding to $n = 2000$) show a similar qualitative behavior. However, the overall level of average absolute deviations is even smaller: Maximum/average values are 0.015/0.008, 0.024/0.011 and 0.021/0.010 for τ, λ^{ℓ} and λ^{u}, respectively.

Results for $d = 10$

We performed the same studies as above for $d = 10$. Since it does not make sense to display the results for all 45 pair-copulas separately, we only report some summary statistics[4]. For a sample size of $n = 200$ the maximum/average deviations were 0.091/0.0483 (τ), 0.069/0.044 (λ^{ℓ}) and 0.068/0.043 (λ^{u}); for $n = 2000$ we observed 0.059/0.018 (τ), 0.037/0.017 (λ^{ℓ}) and 0.037/0.017 (λ^{u}). In comparison to the results for $d = 5$ we detect an increase in deviation, which seems plausible since the dimension of the model increases but the sample sizes are kept constant.

All in all, we see that the sequential fitting of D-vine models to repeated measurement data performs well such that we do not have to hesitate to use it for the real data application in Section 6.

---

[4] The detailed results are of course available on request from the authors.

6 Application

In Section 5 we have seen that our proposed estimation method performs satisfactory. Now we will apply it to real life data. For this purpose, we consider the aortic valve replacement surgery data set `heart.valve` that is taken from the R library `joineR` (Philipson et al., 2017) and has been analyzed in Lim et al. (2008). For this longitudinal study the regression of the left ventricular mass index (LVMI) of $n = 256$ individuals was examined in several follow-up appointments after the surgery, where a new heart valve had been implanted. The total number of examinations is 988 such that the average number of measurements per patient is 3.86, where 10 is the maximum. Table 4 summarizes the sizes of the groups of individuals with exactly j and j or more measurements, respectively, $j = 1, \ldots, 10$.

j	1	2	3	4	5	6	7	8	9	10
patients with j measurements	51	44	41	30	27	15	21	15	6	6
patients with $\geq j$ measurements	256	205	161	120	90	63	48	27	12	6

Table 4: Sizes of the groups of individuals with exactly j and j or more measurements, respectively, $j = 1, \ldots, 10$.

Besides the examination results, for every patient and measurement there are also covariates available. We denote them the way they are stored in the data set `heart.valve`. The following list contains the covariates that we used in our final models as well as a short description, which is basically taken from the documentation of the `joineR` library (Philipson et al., 2017):

- `size`: size of the heart valve in millimeters;

- `sex`: gender of the patient

- `bsa`: body surface area (preoperative)

- `time`: date of measurement (with surgery date as time origin)

The quantity we model is the logarithm of the LMVI. We estimate two different models: a linear mixed model and a D-vine copula based model. We focus on the first five measurements since there are rather few observations for the later measurements. This way we use 832 of the 988 available measurements (84.2%).

Linear mixed model approach
In order to fit a linear mixed model (cf. Section 3.1) to the data we use the function `lme` from the R library `nlme` (Pinheiro et al., 2017). Assuming a homogeneous covariance structure for all individuals i, $i = 1, \ldots, 256$, different correlation structures such as i.i.d. errors, compound symmetry or AR(1) can be selected (cf. Section 3, page 10). We fit a random intercept model, i.e. $Z_i = (1, \ldots, 1)^\top \in \mathbb{R}^{d_i \times 1}$, and compare different (homogeneous) correlation structures for the error terms, namely i.i.d., compound symmetry, AR(1), exponential decay and general (i.e. unrestricted) structure. Note that we perform classical maximum-likelihood estimation (instead of restricted maximum-likelihood estimation, which is often used for linear mixed models) since we need be able to compare the quality of the fit to our D-vine based approach later. The parameter estimates, however, are almost the same for. The best model with respect to

log-likelihood and AIC is the one with the general structure; it contains the covariates `size`, `sex` and `bsa` as well as an `intercept` as fixed effects. The AR(1) error structure, where the (k, ℓ)th entry of Σ_i is given by $\sigma^2 \rho^{|k-\ell|}$, is more parsimonious than the general structure and exhibits a better BIC although log-likelihood and AIC are worse; it contains the covariates `size`, `sex`, `bsa` and `time` as well as an `intercept`. Note that the use of BIC for linear mixed models is controversial (Hedeker and Gibbons, 2006); it is frequently debated which sample size to use for the calculation of BIC (Jones, 2011; Müller et al., 2013; Delattre et al., 2014). In the penalty of the standard BIC all model parameters are weighted with the logarithm of the total number of observations. Here, we used the adjusted BIC for linear mixed models that was developed by Delattre et al. (2014) and that is better comparable to the one we derived for our approach in Proposition 4.1, where each parameter is weighted with the logarithm of the number of observations that directly contribute to its estimation. In the adjusted penalty term of Delattre et al. (2014) the parameters associated with the fixed effects are weighted with the logarithm of the number of measurements and the parameters associated with the random effects are weighted with the logarithm of the number of individuals.

The log-likelihood, AIC and BIC values of the models with the general and the AR(1) structure can be found in Table 5. The remaining structures are not listed there as they performed uniformly worse than the two models.

D-vine copula based approach

As an alternative we will also fit our D-vine based model. As described in Section 2.1 we first deal with the univariate marginal distributions and afterwards estimate the dependence structure. For the marginals we use the univariate marginal regression model that was already estimated for the linear mixed model with AR(1) error correlation structure. Hence, the margins depend on the covariates `size`, `sex`, `bsa` and `time`. In order to transform the measurements to the uniform scale, we apply the estimated normal distribution functions resulting from the regression model (cf. Equation (3.8)). Then a D-vine copula with order 1–2–3–4–5 is fitted to the transformed observations according to the sequential approach from Section 4.2 (using `RVineCopSelect`). In order to avoid unnecessary parameters we apply a Kendall's τ based independence test (significance level $\alpha = 5\%$), which is also implemented in `RVineCopSelect`, to decide for each pair-copula if it is significantly different from an independence copula (for a detailed description see Hollander et al., 2014; Genest and Favre, 2007). The criterion for the selection of the pair-copula families is standard BIC. We fit both a Gaussian D-vine copula, where all pairs are assumed to be bivariate Gaussian, and a general, unrestricted D-vine copula. The result is in both cases a first-order Markov structure, also known as a 1-truncated vine copula, i.e. all pair-copulas in the second, third and fourth tree are the independence. For the Gaussian vine the Kendall's τ values of the Gaussian pairs in the first tree are estimated to be $\hat{\tau}_{1,2} = 0.43$, $\hat{\tau}_{2,3} = 0.54$, $\hat{\tau}_{3,4} = 0.56$ and $\hat{\tau}_{4,5} = 0.61$. For the general D-vine copula the pair-copulas in the first tree are estimated to be the following: $\hat{c}_{1,2} =$ Frank ($\hat{\tau}_{1,2} = 0.49$); $\hat{c}_{2,3} =$ Survival Gumbel ($\hat{\tau}_{2,3} = 0.53$); $\hat{c}_{3,4} =$ Survival Gumbel ($\hat{\tau}_{3,4} = 0.56$); $\hat{c}_{4,5} =$ Frank ($\hat{\tau}_{4,5} = 0.65$). Figure 3 displays pairwise plots of the copula data (transformed to standard normal margins for reasons of comparability) including the contour lines of the corresponding fitted pair-copulas.

We see that there is a positive medium strength of dependence for all pairs in both models (τ-values from 0.43 to 0.61 and 0.49 to 0.65, respectively). The shape of the contours, however, differs considerably between the two models. All four bivari-
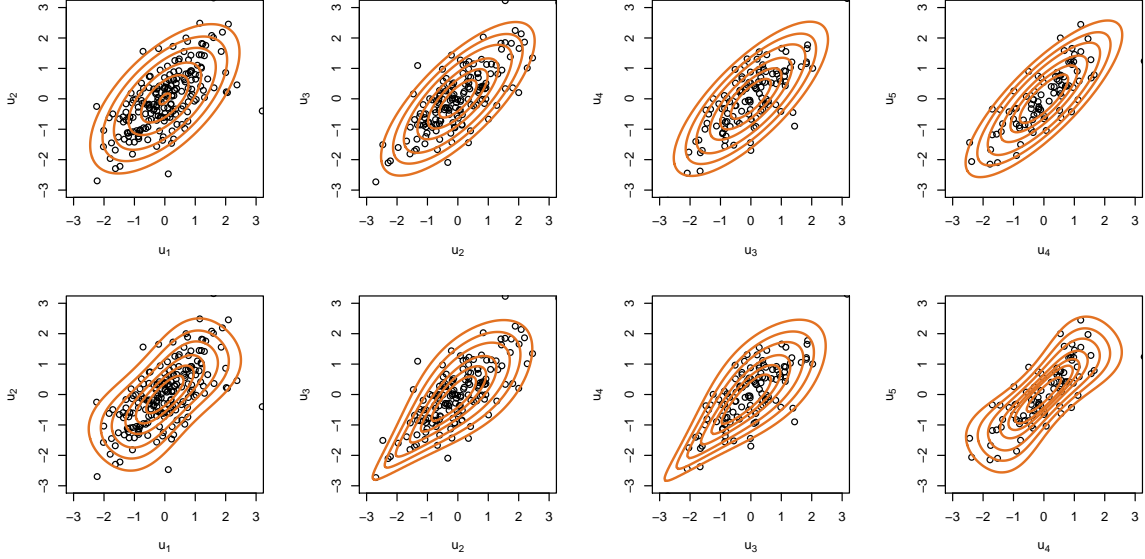
Figure 3: Pairwise plots of the copula data (transformed to standard normal margins) including the contours lines of the fitted pair-copulas of the Gaussian D-vine copula (upper panel) and the general D-vine copula (lower panel), respectively.

ate copulas in the general D-vine model are from different families and non-Gaussian. Whereas $\hat{c}_{1,2}$ and $\hat{c}_{4,5}$ show no tail dependence, the survival Gumbel copula modeling the dependence between the second and the third and the third and the fourth variable exhibits moderate lower tail dependence: $\hat{\lambda}^{\ell}_{2,3} = 0.62$ and $\hat{\lambda}^{\ell}_{3,4} = 0.64$. The fact that the dependence between two consecutive measurements is not constant and non-Gaussian is an indicator that the general D-vine approach might be a better choice than a simple Gaussian dependence model.

Model comparison

In order to see if this is the case we compare the fit of the two D-vine based models (including margins) and the two linear mixed models to the data using the log-likelihood (of the full model including margins), AIC and BIC. Table 5 displays all three model selection criteria. Note that the BIC values of the linear mixed models are calculated as proposed by Delattre et al. (2014) and the ones of the D-vine based models are calculated according to Proposition 4.1.

model	log-likelihood	AIC	BIC	# parameters
general LMM	-99.4	230.9	305.3	16
LMM AR(1)	-108.9	233.8	270.4	8
Gaussian D-vine	-102.7	225.3	265.3	10
general D-vine	$\mathbf{-85.0}$	**190.1**	**230.1**	10

Table 5: Log-likelihood, AIC and BIC values for the fitted linear mixed models with general and AR(1) structure and the Gaussian and the general D-vine based models (including margins). Bold values indicate the best model fit according to the respective model selection criteria.

One can see that the unrestricted D-vine model performs uniformly better than the Gaussian one. This is a clear indicator that the normality assumption for the dependence is not really suited. Nevertheless, due to their flexibility, both D-vine

based models yield a considerably better fit than the two linear mixed model with respect to log-likelihood, AIC and BIC; only the log-likelihood of the Gaussian D-vine is slightly worse than the one of the general LMM. We see that the D-vine based approaches are able to capture the structure of the data better since the flexibility of the D-vine helps to fit the dependence structure more appropriately. This is especially important if the deviation from Gaussianity is strong.

Quantile prediction

As a final application we illustrate how conditional quantiles for the $(j + 1)$st measurement of an individual with j measurements can be determined using the general D-vine copula based approach and the linear mixed model with AR(1) error structure. For this purpose, we select three representative individuals with $d_i = 4$ measurements from the data set such that they have had rather low ($\mathbf{y}^1 = (4.63, 4.62, 4.66, 4.91)^\top$), medium ($\mathbf{y}^2 = (5.26, 5.13, 5.00, 5.19)^\top$) and high ($\mathbf{y}^3 = (5.90, 5.80, 5.67, 5.31)^\top$) measurement values so far, respectively. We will use the corresponding observed covariate values of $\mathbf{x}^i = (x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})^\top = (\texttt{size}_i, \texttt{sex}_i, \texttt{bsa}_i, \texttt{time}_i)^\top$ that are given by $\mathbf{x}^1 = (29, 0, 1.93, 3.15)^\top$, $\mathbf{x}^2 = (25, 0, 1.65, 5.48)^\top$ and $\mathbf{x}^3 = (25, 0, 1.71, 3.19)^\top$, respectively.

We pretend that the three selected individuals have only had three measurements so far. Then we predict the median, i.e. the 50% quantile, and a 90% confidence interval, i.e. the 5% and the 95% quantile, of the fourth measurement based on the three measurements y_1^i, y_2^i, y_3^i for both models and compare the results to the true value of the fourth measurement.

For the linear mixed model we know that the joint distribution of the measurements of one individual is a multivariate normal distribution with mean and variance as in Equation (3.10). Having estimated the corresponding parameters we can easily determine the conditional distribution of the fourth measurement given the first three measurements, which is given by a univariate normal distribution (see for example Joe, 2014, Section 2.6). The quantiles of univariate normal distributions are known such that we can easily compute the desired quantities. The estimated median for individual 1 is 4.78, the estimated 90% confidence interval is given by $(4.37, 5.18)$. For individuals 2 and 3 the medians are 5.21 and 5.73 and the confidence intervals are $(4.81, 5.61)$ and $(5.32, 6.13)$, respectively. We see that the observed value of the fourth measurement of individuals 1 and 2 are inside the confidence bounds; for the third individual the confidence interval does not contain the observed measurement value: $5.31 \notin (5.32, 6.13)$. Note that due to the normality of the conditional distribution the confidence bounds are symmetric around the mean (which is also the median).

For comparison we determine the conditional quantiles using the general D-vine based model. Since we want to apply Equation (2.6), which uses the inverse conditional distribution function on the copula level $C_{4|1:3}^{-1}(\,\cdot\,|F_1^i(y_1^i), F_2^i(y_2^i), F_3^i(y_3^i))$ and the inverse of the marginal distribution function of the fourth measurement $(F_4^i(\cdot))^{-1}$, we first of all transform the measurements 1, 2 and 3 to the copula scale by $\hat{u}_j^i := \hat{F}_j^i(y_j^i)$, where $\hat{F}_j^i$ are the marginal linear regression estimates obtained from the linear mixed model. For each individual $i = 1, 2, 3$ and $\alpha \in \{0.05, 0.50, 0.95\}$ we use the estimated general D-vine copula to calculate $\hat{q}_u^i(\alpha) := \hat{C}_{4|1:3}^{-1}(\alpha|u_1^i, \ldots, u_3^i)$. Since $\hat{F}_4^i(\cdot)$ is again a univariate normal distribution we can easily determine its inverse and apply $(\hat{F}_4^i(\alpha))^{-1}$ to $\hat{q}_u^i(\alpha)$ to obtain the conditional quantiles $\hat{q}_y^i(\alpha) = (\hat{F}_4^i(\alpha))^{-1}(\hat{q}_u^i(\alpha))$. The estimated median for individual 1 is 4.65, the estimated 90% confidence interval is given by $(4.42, 5.02)$. For individuals 2 and 3 the medians are 5.18 and 5.60 and the confidence intervals are

$(4.82, 5.59)$ and $(5.13, 6.03)$, respectively. Thus all three confidence intervals contain the corresponding observed measurement values. Note that the 5% and 95% quantiles are in general not symmetric around the 50% quantile in this case.

In practice it might be interesting to investigate the influence of covariates on the estimated conditional quantiles. As an example we illustrate how the quantiles depend on the variable `bsa`. In Figure 4 we show the resulting estimates for the median (solid lines) and the confidence intervals (dashed lines) of individuals 1 (left), 2 (middle) and 3 (right) depending on the size of the heart valve for the linear mixed model (gray lines) and the D-vine based model (black lines). The estimated quantiles for the actual covariate specifications of our three selected individuals are given by the intersections dotted vertical lines (indicating the true `bsa` values) and the quantiles lines; for example the median is marked at the intersection point of the confidence interval and the corresponding median line. The actual observed measurement value y_4^i is also added as a circle.
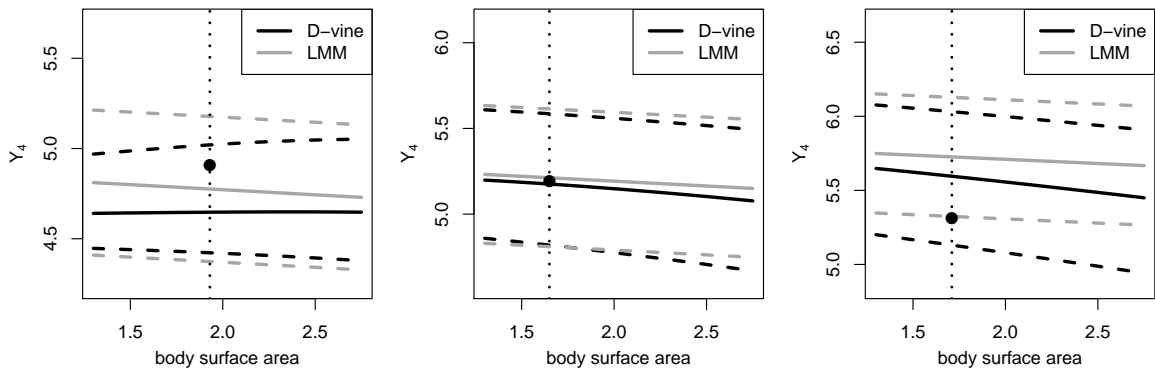
Figure 4: Estimated median (solid lines) and confidence intervals (dashed lines) of the fourth measurement for individuals 1 (left column), 2 (middle column) and 3 (right column) depending on the size of the heart valve for the LMM (gray) and the D-vine based model (black). Vertical lines indicate the true `bsa` values. Actual observed values are marked as circles.

First of all, we see that the quantiles estimated from the linear mixed model depend linearly on `bsa`. This is clear since the covariates only influence the estimated mean of the distribution of the fourth measurement given the first three. This influence is the same for all quantiles, i.e. the slope of the all gray lines is the same. Since the standard deviation does not depend on the covariate values all confidence intervals have the same width (even for different individuals). Normality implies that the confidence bounds lie symmetric around the mean (median).

The quantiles estimated on the basis of the D-vine based model, however, inherit the flexibility of the D-vine model and do not depend linearly on the `bsa` value. The width of the confidence intervals varies among the three individuals and even for one individual depending on the covariate values. The slope for different quantiles can even be positive and negative for one individual. These phenomena can be seen very clearly in left plot in Figure 4) corresponding to the first individual.

This application illustrated how easily such investigations can be performed with both the linear mixed model and the D-vine based model. Analyzing the results, however, has made clear that the flexibility of the D-vine based model is a non-ignorable advantage over the linear mixed model. For both models it is eye-catching that the differences between the three individuals are considerable. This shows how important it can be to use the information available when making predictions for the future.

7   Conclusion and outlook

This article presented an intuitive and easily interpretable D-vine copula based model with arbitrary margins for (possibly) unbalanced longitudinal data. The model was compared to linear mixed models and proved to be a generalization of this model class under the assumption that the correlation structure was homogeneous over the individuals. Further, we developed a BIC adjustment for our model. Being based on D-vine copulas our proposed model benefited from the possibility to model the underlying dependence structure very flexibly. Since we did not impose any restrictions on the univariate marginal distributions, this adds even more flexibility to the model. As joint estimation of the D-vine copula would become rather slow in high dimensions, we proposed a fast sequential alternative, where even missing data values could be handled without causing problems. Due to the nested nature of D-vine models our approach further easily allowed for predicting future events. In the application to the heart surgery data set the proposed model was able to fit the data considerably better than the linear mixed models. If data exhibited an even more complicated dependence structure than the considered data set (possibly including stronger tail dependence, asymmetries etc.), the Gaussian assumption of linear mixed models would certainly be so strongly violated that changing to a more flexible model would be inevitable.

In an ongoing research project the D-vine based modeling approach is extended to time-to-event data with right-censoring (Barthel et al., 2017).

Acknowledgments

Appendix

A   Proof of Proposition 4.1

We will prove the statement of Proposition 4.1 for $d = 2$ in order to present the basic idea. The extension to higher dimensions works similarly but involves more tedious calculations. In our proof we adapt the derivation from Neath and Cavanaugh (2012). Since our proof is very similar up to the last step, we refer the reader to their paper for a more detailed argumentation.

BIC is used for model selection when different parametric candidate models $M_1, \ldots, M_K$ are available to describe a data set $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}$. Further, let $L(\boldsymbol{\theta}_k | \mathbf{Y})$ be the likelihood corresponding to model M_k, depending on the parameters $\boldsymbol{\theta}_k \in \Omega_k$, where $\Omega_k \subseteq \mathbb{R}^{p_k}$ is the space of admissible parameters. Let $\pi(k)$ be the prior probability corresponding to model M_k and $g(\boldsymbol{\theta}_k | k)$ denote a prior on $\boldsymbol{\theta}_k$ given the model M_k. Using Bayes' Theorem we obtain the joint posterior of M_k and $\boldsymbol{\theta}_k$:

$$h(k, \boldsymbol{\theta}_k | \mathcal{Y}) = \frac{\pi(k) g(\boldsymbol{\theta}_k | k) L(\boldsymbol{\theta}_k | \mathcal{Y})}{m(\mathcal{Y})},$$

where $m(\mathcal{Y})$ denotes the marginal distribution of $\mathbf{Y}$. We are interested in finding the highest posterior probability of M_k given $\mathcal{Y}$, which can be expressed as

$$P(k|\mathcal{Y}) = \frac{\pi(k)}{m(\mathcal{Y})} \int_{\Omega_k} L(\boldsymbol{\theta}_k|\mathcal{Y}) g(\boldsymbol{\theta}_k|k) \mathrm{d}\boldsymbol{\theta}_k.$$

Since maximizing $P(k|\mathcal{Y})$ is equivalent to minimizing $-2 \log P(k|\mathcal{Y})$ with respect to k and $m(\mathcal{Y})$ does not depend on k, we will from now on consider

$$S(k|\mathcal{Y}) := -2 \log \pi(k) - 2 \log \int_{\Omega_k} L(\boldsymbol{\theta}_k|\mathcal{Y}) g(\boldsymbol{\theta}_k|k) \mathrm{d}\boldsymbol{\theta}_k. \tag{A.11}$$

In order to be able to approximate the integrand from Equation (A.11) we perform a second-order Taylor series expansion of the log-likelihood $\log L(\boldsymbol{\theta}_k|\mathcal{Y})$ around the maximum-likelihood parameter estimate $\hat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta}_k \in \Omega_k} L(\boldsymbol{\theta}_k|\mathcal{Y})$:

$$\log L(\boldsymbol{\theta}_k|\mathcal{Y}) \approx \log L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^\top \frac{\partial \log L(\boldsymbol{\theta}_k|\mathcal{Y})}{\partial \boldsymbol{\theta}_k} \bigg|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}$$

$$+ \frac{1}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^\top \left[\frac{\partial^2 \log L(\boldsymbol{\theta}_k|\mathcal{Y})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^\top} \bigg|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k} \right] (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)$$

Since $\hat{\boldsymbol{\theta}}_k$ maximizes $L(\boldsymbol{\theta}_k|\mathcal{Y})$, and hence also $\log L(\boldsymbol{\theta}_k|\mathcal{Y})$, we obtain

$$L(\boldsymbol{\theta}_k|\mathcal{Y}) \approx L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^\top H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) \right\}$$

where we denote the negative Hessian matrix of the log-likelihood by

$$H(\boldsymbol{\theta}_k|\mathcal{Y}) := -\frac{\partial^2 \log L(\boldsymbol{\theta}_k|\mathcal{Y})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^\top}.$$

Neath and Cavanaugh (2012) and Cavanaugh and Neath (1999) argue that the above approximations hold for large samples $\mathcal{Y}$ and further justify the use of a non-informative prior $g(\boldsymbol{\theta}_k|k) = 1$ for any $\boldsymbol{\theta}_k \in \Omega_k$. Thus,

$$L(\boldsymbol{\theta}_k|\mathcal{Y}) \approx L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})(2\pi)^{p_k/2} \left| H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) \right|^{-1/2}. \tag{A.12}$$

Plugging Equation (A.12) into Equation (A.11) yields

$$S(k|\mathcal{Y}) \approx -2 \log \pi(k) - 2 \log L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) - p_k \log \pi + \log \left| H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) \right|. \tag{A.13}$$

In order to compute the determinant of $H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})$ we consider the (ℓ, m)th entry $H_{\ell,m}$ of $H(\boldsymbol{\theta}_k|\mathcal{Y})$. Since $d = 2$ the parameter vector $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_k^1, \boldsymbol{\theta}_k^2, \boldsymbol{\theta}_k^3)^\top$ can be split up such that $\boldsymbol{\theta}_k^j \in \mathbb{R}^{q_j}$ parametrize the marginal distributions F_j of the jth measurement, $j = 1, 2$ and $\boldsymbol{\theta}_k^3 \in \mathbb{R}^{q_3}$ is the parameter vector of the copula $c_{1,2}$ with $p_k = q_1 + q_2 + q_3$. For the sake of notation we assume that $\mathcal{Y}$ is ordered such that $\mathcal{Y}^2 = \{\mathbf{y}^1, \ldots, \mathbf{y}^{n_2}\}$ and

$\mathcal{Y}^1 = \{\mathbf{y}^{n_2+1}, \dots, \mathbf{y}^n\}$ and further recall that $N_1 = n_1 + n_2 = n$ and $N_2 = n_2$. We have

$$H_{\ell,m} = -\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\sum_{i=1}^{n}\log L(\boldsymbol{\theta}_k|\mathbf{y}^i)$$

$$= -\sum_{i=1}^{N_1}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log f_1(y_1^i|\boldsymbol{\theta}_k^1) - \sum_{i=1}^{N_2}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log f_2(y_2^i|\boldsymbol{\theta}_k^2)$$

$$- \sum_{i=1}^{N_2}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log c_{1,2}(F_1(y_1^i|\boldsymbol{\theta}_k^1), F_2(y_2^i|\boldsymbol{\theta}_k^2)|\boldsymbol{\theta}_k^3)$$

$$= N_1\left[-\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log f_1(y_1^i|\boldsymbol{\theta}_k^1)\right] + N_2\left[-\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log f_2(y_2^i|\boldsymbol{\theta}_k^2)\right]$$

$$+ N_2\left[-\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{\partial^2}{\partial\theta_\ell\partial\theta_m}\log c_{1,2}(F_1(y_1^i|\boldsymbol{\theta}_k^1), F_2(y_2^i|\boldsymbol{\theta}_k^2)|\boldsymbol{\theta}_k^3)\right]$$

Assuming that the data set is large, i.e. N_1 and N_2 are large, the expressions in the brackets (approximately) represent entries of the Fisher information matrices

$$I_1 = I_1(\boldsymbol{\theta}_k^1|\mathcal{Y}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_k^1\partial(\boldsymbol{\theta}_k^1)^\top}\log f_1(Y_1|\boldsymbol{\theta}_k^1)\right] \in \mathbb{R}^{q_1\times q_1},$$

$$I_2 = I_2(\boldsymbol{\theta}_k^2|\mathcal{Y}^2) = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_k^2\partial(\boldsymbol{\theta}_k^2)^\top}\log f_2(Y_2|\boldsymbol{\theta}_k^2)\right] \in \mathbb{R}^{q_2\times q_2}$$

and

$$I_3 = \begin{pmatrix} I_3^{1,1} & I_3^{1,2} & I_3^{1,3} \\ I_3^{2,1} & I_3^{2,2} & I_3^{2,3} \\ I_3^{3,1} & I_3^{3,2} & I_3^{3,3} \end{pmatrix} = I_3((\boldsymbol{\theta}_k^1, \boldsymbol{\theta}_k^2, \boldsymbol{\theta}_k^3)|\mathcal{Y}^2) \in \mathbb{R}^{(q_1+q_2+q_3)\times(q_1+q_2+q_3)},$$

where

$$I_3^{\ell,m} = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_k^\ell\partial(\boldsymbol{\theta}_k^m)^\top}\log c_{1,2}(F_1(Y_1|\boldsymbol{\theta}_k^1), F_2(Y_2|\boldsymbol{\theta}_k^2)|\boldsymbol{\theta}_k^3)\right] \in \mathbb{R}^{q_\ell\times q_m}.$$

Thus, $H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})$ can be written as

$$H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) = \begin{pmatrix} N_1 I_1 + N_2 I_3^{1,1} & N_2 I_3^{1,2} & N_2 I_3^{1,3} \\ N_2 I_3^{2,1} & N_2 I_2 + N_2 I_3^{2,2} & N_2 I_3^{2,3} \\ N_2 I_3^{3,1} & N_2 I_3^{3,2} & N_2 I_3^{3,3} \end{pmatrix}.$$

Using the formula for the determinant of block-matrices (Silvester, 2000) we obtain

$$\left|H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})\right| = N_1^{q_1} N_2^{q_2+q_3}\left|I_1 + \frac{N_2}{N_1}I_3^{1,2} - I_3^{1,3}(I_3^{3,3})^{-1}I_3^{3,1} + \frac{N_2}{N_1}\left[I_3^{1,2} - I_3^{1,3}(I_3^{3,3})^{-1}I_3^{3,2}\right]\right.$$

$$\times \left[I_2 + I_3^{2,2} - I_3^{2,3}(I_3^{3,3})^{-1}I_3^{3,2}\right]^{-1}\left[I_3^{2,1} - I_3^{2,3}(I_3^{3,3})^{-1}I_3^{3,1}\right]\Bigg|$$

$$\times \left|I_2 + I_3^{2,2} - I_3^{2,3}(I_3^{3,3})^{-1}I_3^{3,2}\right|\left|I_3^{3,3}\right|$$

$$=: N_1^{q_1} N_2^{q_2+q_3}\, a(N_1, N_2).$$

Note that since N_2/N_1 is bounded between 0 and 1, $a(N_1, N_2)$ is also bounded. Plugging the expression for $\left|H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})\right|$ into Equation (A.13) we obtain

$$S(k|\mathcal{Y}) \approx -2\log\pi(k) - 2\log L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) - p_k\log\pi + q_1\log N_1 + (q_2+q_3)\log N_2 + \log a(N_1, N_2).$$

Discarding the terms that are bounded as the sample size goes to infinity yields

$$S(k|\mathcal{Y}) \approx -2\log L(\hat{\boldsymbol{\theta}}_k|\mathcal{Y}) + \Delta p_1 \log N_1 + \Delta p_2 \log N_2$$

since $\Delta p_1 = q_1$ and $\Delta p_2 = q_2 + q_3$. This proves the statement for $d = 2$. The proof of Proposition 4.1 in higher dimensions only differs from the above in that the calculations necessary to compute the determinant of $H(\hat{\boldsymbol{\theta}}_k|\mathcal{Y})$ are much more involved since one has to compute the determinant of a $(d(d+1)/2) \times (d(d+1)/2)$ block matrix.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics*, 44:182–198.

Acar, E. F., Genest, C., and Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90.

Airy, G. (1861). *On the algebraic and numerical theory of errors of observations and the combination of observations*. London: Macmillan.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. New York, NY: Springer.

Barthel, N., Geerdens, C., Killiches, M., Janssen, P., and Czado, C. (2017). Vine models for recurrent event times subject to right censoring. *Working paper*.

Bedford, T. and Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068.

Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics-Theory and Methods*, 28(1):49–66.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, pages 93–109. Springer Berlin Heidelberg, Berlin, Heidelberg.

Czado, C., Gärtner, F., and Min, A. (2011). Analysis of Australian electricity loads using joint Bayesian inference of D-vines with autoregressive margins. *Dependence Modeling: Vine Copula Handbook, World Scientific Publishing*, pages 265–280.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer Science & Business Media.

Delattre, M., Lavielle, M., Poursat, M.-A., et al. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, 8(1):456–475.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press.

Diggle, P. J. and Donnelly, J. B. (1989). A selected bibliography on the analysis of repeated measurements and related areas. *Australian Journal of Statistics*, 31(1):183–193.

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: models, methods and applications*. Springer Science & Business Media.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*, volume 451. John Wiley & Sons.

Hobæk Haff, I., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction — Simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296–1310.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric statistical methods*. John Wiley & Sons.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189.

Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25):3050–3056.

Killiches, M., Kraus, D., and Czado, C. (2017). Examination and visualisation of the simplifying assumption for vine copulas in three dimensions. *Australian & New Zealand Journal of Statistics*, 59(1):95–117.

Kim, D., Kim, J.-M., Liao, S.-M., and Jung, Y.-S. (2013). Mixture of D-vine copulas for modeling dependence. *Computational Statistics & Data Analysis*, 64:1–19.

Kraus, D. and Czado, C. (2017a). D-vine copula based quantile regression. *Computational Statistics and Data Analysis*, 110C:1–18.

Kraus, D. and Czado, C. (2017b). Growing simplified vine copula trees: improving Dißmann's algorithm. *arXiv preprint arXiv:1703.05203*.

Lambert, P. and Vandenhende, F. (2002). A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21(21):3197–3217.

Lim, E., Ali, A., Theodorou, P., Sousa, I., Ashrafian, H., Chamageorgakis, T., Duncan, A., Henein, M., Diggle, P., and Pepper, J. (2008). Longitudinal study of the profile and predictors of left ventricular mass regression after stentless aortic valve replacement. *The Annals of Thoracic Surgery*, 85(6):2026–2029.

Lindsey, J. and Lindsey, P. (2006). Multivariate distributions with correlation matrices for nonlinear repeated measurements. *Computational Statistics & Data Analysis*, 50(3):720–732.

Lindsey, J. K. (1999). *Models for repeated measurements*. Oxford University Press.

Lorch, R. F. and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):149.

Ludbrook, J. (1994). Repeated measurements and multiple comparisons in cardiovascular research. *Cardiovascular Research*, 28(3):303–311.

Malin, A. and Linnakylä, P. (2001). Multilevel modelling in repeated measures of the quality of finnish school life. *Scandinavian Journal of Educational Research*, 45(2):145–166.

McCulloch, C., Searle, S., and Neuhaus, J. (2011). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. Wiley.

Meester, S. G. and MacKay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics*, 50(4):954–963.

Müller, S., Scealy, J. L., Welsh, A. H., et al. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.

Nai Ruscone, M. and Osmetti, S. A. (2017). Modelling the dependence in multivariate longitudinal data by pair copula decomposition. In Ferraro, M. B. e. a., editor, *Soft Methods for Data Science*, pages 373–380. Springer International Publishing, Cham.

Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203.

Newbold, E. M. (1927). Practical applications of the statistics of repeated events' particularly to industrial accidents. *Journal of the Royal Statistical Society*, 90(3):487–547.

Philipson, P., Sousa, I., Diggle, P. J., Williamson, P., Kolamunnage-Dona, R., Henderson, R., and Hickey, G. L. (2017). *joineR: Joint Modelling of Repeated Measurements and Time-to-Event Data*. R package version 1.1.0.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

Potvin, C., Lechowicz, M. J., and Tardif, S. (1990). The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. *Ecology*, 71(4):1389–1400.

Ren, X., Li, S., Lv, C., and Zhang, Z. (2014). Sequential dependence modeling using Bayesian theory and D-vine copula and its application on chemical process risk prediction. *Industrial & Engineering Chemistry Research*, 53(38):14788–14801.

Schepsmeier, U., Stöber, J., Brechmann, E. C., Gräler, B., Nagler, T., and Erhardt, T. (2017). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.1.1.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shen, C. and Weissfeld, L. (2006). A copula model for repeated measurements with non-ignorable non-monotone missing outcome. *Statistics in Medicine*, 25(14):2427–2440.

Shi, P., Feng, X., Boucher, J.-P., et al. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, 10(2):834–863.

Shi, P. and Yang, L. (2016). Pair copula constructions for insurance experience rating. *Working paper*.

Silvester, J. R. (2000). Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467.

Sklar, A. (1959). Fonctions dé Repartition á n Dimensions et leurs Marges. *Publ. Inst. Stat. Univ. Paris*, 8:229–231.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, 105(492):1467–1479.

Smith, M. S. (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting*, 31(3):815–833.

Spanhel, F. and Kurz, M. S. (2015). Simplified vine copula models: Approximations based on the simplifying assumption. *arXiv preprint arXiv:1510.06971*.

Stöber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions—limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118.

Sun, J., Frees, E. W., and Rosenberg, M. A. (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics*, 42(2):817–830.

Vatter, T. and Nagler, T. (2016). Generalized additive models for pair-copula constructions. *Available at SSRN 2817949*.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34.