Biometrics 000, 1–26 DOI: 000
000 0000

A time-heterogeneous D-vine copula model for unbalanced and unequally spaced longitudinal data

Md Erfanul Hoque[1,5], Elif F. Acar[1,2,3,*], and Mahmoud Torabi[1,4]

[1]Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada.

[2]The Hospital for Sick Children, Toronto, ON, M5G 1X8, Canada.

[3]Department of Statistical Sciences, University of Toronto, Toronto, ON, M5S 3G3, Canada.

[4]Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, R3E 0W3, Canada.

[5]Department of Statistics, University of Dhaka, Dhaka 1000, Bangladesh.

email: elif.acar@umanitoba.ca

Summary: In many longitudinal studies, the number and timing of measurements differ across study subjects. Statistical analysis of such data requires accounting for both the unbalanced study design and the unequal spacing of repeated measurements. This paper proposes a time-heterogeneous D-vine copula model that allows for time adjustment in the dependence structure of unequally spaced and potentially unbalanced longitudinal data. The proposed approach not only offers flexibility over its time-homogeneous counterparts but also allows for parsimonious model specifications at the tree or vine level for a given D-vine structure. It further provides a robust strategy to specify the joint distribution of non-Gaussian longitudinal data. The performance of the time-heterogeneous D-vine copula models are evaluated through simulation studies and by a real data application. Our findings suggest improved predictive performance of the proposed approach over the linear mixed-effects model and time-homogeneous D-vine copula model.

Key words: Copula; D-vine; Longitudinal study; Missing data; Time-heterogeneity.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Longitudinal studies collect repeated measurements from subjects over time with an aim to understand the dependence mechanisms among these measurements. In many cases, the number of measurements differs across study subjects due to dropouts or failure to follow-up so that the data is unbalanced. In addition, the timing of measurements may differ both within and across study subjects due to lack of or noncompliance with scheduled visit times, which makes the data unequally spaced over time. Statistical analysis of such data requires accounting for both the unbalanced study design and the spacing of repeated measurements.

There have been many developments to analyze longitudinal data. While the linear mixed-effects model (LMM) constitutes the most popular approach to handle continuous longitudinal data outcomes, several other approaches have been proposed to relax distributional assumptions and to handle non-Gaussian longitudinal data. An extensive review of different approaches can be found, for instance, in Diggle (2002), Molenberghs and Verbeke (2005) and Fitzmaurice et al. (2008).

Despite the vast literature on longitudinal data analysis, comparatively few studies have considered modeling strategies for longitudinal data that are both unbalanced and unequally spaced. Rosner and Muñoz (1988) presented a non-linear regression method for autoregressive modeling of longitudinal data with unequally spaced and/or missing observations. Their approach included a lagged value of the outcome as an independent variable and assumed the within-subject errors to be independent after conditioning on the lagged outcome. Jones and Ackerson (1990) and Jones and Boadi-Boateng (1991) considered, respectively, a continuous-time autoregressive moving average structure and a continuous-time autoregressive structure with observational error for modeling the within-subject covariance structure in longitudinal data observed at different unequally spaced time points. Nunez-Anton and Woodworth (1994) proposed a three-parameter covariance model with a power transformation of time scale

to analyze irregularly spaced longitudinal data. All these approaches assumed a Gaussian error structure. Following a generalized estimating equations (GEE) based approach, Shults and Chaganty (1998) proposed the use of quasi-least squares to estimate and select a suitable working correlation structure for unbalanced and unequally spaced longitudinal data. They considered the first-order autoregressive, Markov and generalized Markov correlation structures for the specification of the within-subject working correlation structure.

Copulas offer a more general and versatile framework to model dependence than correlation based models. They have been used to model within-subject dependence in longitudinal data in medical research (Lambert and Vandenhende, 2002), healthcare records (Sun et al., 2008), insurance pricing (Frees and Wang, 2006) and energy consumption (Smith et al., 2010). These works exclusively focused on the balanced setting and did not address any discrepancy in the frequency and timing of measurements. The problem of modeling unbalanced longitudinal data has been recently tackled in Shi et al. (2016) using a Gaussian copula model and in Killiches and Czado (2018) using a D-vine copula model. Both models assumed a homogeneous dependence structure for all subjects. This assumption, however, is not realistic for unequally spaced longitudinal data, whether balanced or unbalanced, as measurements taken close in time are likely to be more dependent than measurements taken far apart. Hence, time intervals between measurements should be taken into account in the modeling strategy.

Here, we propose a time-heterogeneous D-vine copula model that allows for time adjustment in the dependence structure of unequally spaced and potentially unbalanced longitudinal data. Vine copulas are multivariate dependence models constructed sequentially from bivariate building blocks, called pair copulas (Joe, 1997; Bedford and Cooke, 2002), with D-vine copulas being a special class particularly suited for modeling serial dependence (Smith et al., 2010). Since these models consist of a series of bivariate copulas, time heterogeneity can be included by allowing each pair copula to depend on the length of the time interval between

the corresponding measurements. Specifically, we introduce a two-parameter exponential decay model for the pair copula dependence parameters that incorporates the time interval between the two measurements. To facilitate inference for unbalanced longitudinal data, we consider a monotone missing data pattern, which typically occurs when subjects drop out from the study and assume that data are missing completely at random.

The merits of our proposed approach are threefold. It offers flexibility over the time-homogeneous D-vine copula model of Killiches and Czado (2018) by accounting for both visit to visit and subject to subject variation in the spacing between measurements. Owing to its added flexibility, the proposed approach allows parsimonious model specifications at the tree or vine level for a given D-vine structure, which would be advantageous in high-dimensional problems. It further facilitates robust specification of the joint distribution of longitudinal data in situations where the multivariate Gaussian assumption is questionable. We outline estimation methods for the time-heterogeneous D-vine copula models and discuss parsimonious model specifications. The proposed approach is evaluated in extensive simulations and demonstrated using a subset of the Manitoba Follow-up Study (MFUS). Our findings demonstrate the gain in predictive performance achieved by using our proposed approach over the LMM and the time-homogeneous D-vine copula model.

The paper is organized as follows. Section 2 introduces our motivating data example from the MFUS. Section 3 presents the proposed time-heterogeneous D-vine copula model and the estimation methods. Section 4 contains simulation studies which compare the performance of the time-heterogeneous D-vine copula model with those of the time-homogeneous D-vine copula model and LMM. Section 5 illustrates the proposed approach through analysis of the MFUS data. Concluding remarks are given in Section 6. Additional results from the simulations and data analysis are collected in the Supporting Information available at *Biometrics* online.

2. Motivating Example

Our motivating data example comes from the MFUS which is the largest and longest running investigation of cardiovascular disease in Canada. The study was established at the University of Manitoba, Canada, on July 1, 1948 with a cohort that consists of 3983 men who were recruited in the Royal Canadian Air Force during World War II (Mathewson et al., 1965). The mean age in the cohort was around 31 years, with about 90% between age 20 and 39 years. The baseline measurements of diastolic blood pressure (DBP) and body mass index (BMI) (mean $\pm$ standard deviation) were reported as $121 \pm 10\,\text{mmHg}$ and $23.8 \pm 2.7\text{kg/m}^2$, respectively (Tate et al., 2014). Even though routine medical examinations were requested at regular intervals and efforts were made to conduct these examinations on scheduled times, subjects often missed assessments or showed up between scheduled visits. Hence, the number and timing of the measurements were quite different within and across subjects.

High blood pressure is one of the most important risk factors for cardiovascular disease. It is also a risk factor for myocardial infarction, stroke, congestive heart failure, and peripheral vascular disease (Whelton, 1994; MacMahon et al., 1990; Flebach et al., 1989). It is known that being overweight and obesity are associated with high blood pressure (World Health Organization, 2002). Hence, it is of interest to study the longitudinal elevation of DBP and its risk factors including age, BMI and IHD (ischemic heart disease) status (see Section 5 for more details).

Figure 1 displays the DBP measurements collected from five randomly selected subjects from the MFUS. As can be seen, subjects have different number of visits reflecting the unbalancedness of the data. Most existing approaches, including the LMM and time-homogeneous D-vine copula model, consider only the visit label when modeling the dependence among the repeated measurements, hence view the data as in the left panel. However, the timing of measurements, as quantified by the age at visits in the right panel, may also provide

important information, especially when subjects show a considerable variation in the spacing of their measurements. For instance, Subject 2 and Subject 3 both have four measurements, however, being taken in a shorter time frame, the first two measurements of Subject 2 are likely to be more strongly associated than the first two measurements of Subject 3. Hence, it may not be reasonable to assume a common dependence structure for the subjects with the same number of measurements.

[Figure 1 about here.]

In addition, assuming the joint distribution of the repeated measurement to be multivariate Gaussian may not always be warranted or is difficult to justify. For instance, the marginal distributions of the DBP measurements across the first five visits show only a slight divergence from normality, hence they may fulfill the Gaussian assumption (see Web Figure 2 in Web Appendix C of the Supporting Information). However, the same is difficult to claim for their dependence structure and joint distribution since some pairs show evidence of asymmetric dependence (e.g., stronger upper tail dependence in pairs of consecutive measurements). Hence, the LMM framework may not be appropriate to analyze this dataset.

3. Time-heterogeneous D-vine Copula Model

3.1 *Model*

Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be the random vector representing d continuous repeated measurements from one subject observed at time points $t_1, t_2, \ldots, t_d$ with $t_1 < t_2 < \ldots < t_d$. Then, by Sklar's theorem (Sklar, 1959), the joint distribution F of $\mathbf{Y}$ given $t_1, \ldots, t_d$ can be written as

$$F(y_1, \ldots, y_d \mid t_1, \ldots, t_d) = C(F_1(y_1 \mid t_1), \ldots, F_d(y_d \mid t_d) \mid t_1, \ldots, t_d), \tag{1}$$

where F_j is the conditional marginal distribution of Y_j given t_j for $j = 1, \ldots, d$ and C is the d-dimensional (conditional) copula, which gives the dependence among the repeated

measurements conditional on the observed time points. We assume that the visit times are continuous and independent of longitudinal outcomes as well as past visit times.

For longitudinal data, the repeated measurements exhibit a clear temporal ordering, hence the use of D-vine copulas with order 1-2-$\cdots$-d has been suggested (Smith et al., 2010; Killiches and Czado, 2018). Therefore, the copula C can be defined using the D-vine decomposition with pair copulas of consecutive measurements in the first tree (see Figure 2).

Let $U_j \equiv F_j(y_j \mid t_j)$, and $c(u_1, \ldots, u_d) = \partial^d C(u_1, \ldots, u_d)/\partial u_1 \cdots \partial u_d$ represent the copula density. Then, under the D-vine model, we can decompose copula density c as

$$c(u_1, \ldots, u_d \mid t_1, \ldots, t_d) = \prod_{k=1}^{d-1} \prod_{j=1}^{d-k} c_{j(j+k);S_{jk}} \left(u_{j|S_{jk}}, u_{j+k|S_{jk}} \mid t_j, t_{j+k}\right), \tag{2}$$

where $S_{jk} = \{j+1, \ldots, j+k-1\}$ is the conditioning set and $c_{j(j+k);S_{jk}}$ is the (conditional) pair copula of (U_j, U_{j+k}) given $\boldsymbol{U}_{S_{jk}} = (U_{j+1}, \ldots, U_{j+k-1})$. The conditional variables $U_{j|S_{jk}}$ and $U_{j+k|S_{jk}}$ are obtained sequentially using

$$\frac{\partial C_{ml|S_{-l}}\left(u_{m|S_{-l}}, u_{l|S_{-l}}\right)}{\partial u_{l|S_{-l}}}, \tag{3}$$

with $(m, l) = (j, j+k-1)$ and $(m, l) = (j+k, j+1)$, respectively, where S_{-l} denotes the index set excluding the l^{th} component.

Note that a common assumption to build the vine copula models is that the conditioning variables do not affect the conditional pair copulas appearing in higher level trees. We also make this so-called simplifying assumption here. However, we allow the pair copulas in Equation (2) to depend on the observed time points, and hence, incorporate potential time-heterogeneity in the dependence structure. Modeling the general functional form of $c(\cdot, \cdot \mid t_j, t_{j+k})$ can be quite challenging. Here, we consider bivariate parametric copula families $\mathcal{C} = \{c : \theta \in \Theta\}$ where the copula parameter θ varies with the length of time interval between the two measurements rather than the actual time points, i.e., $\theta_{j(j+k)}(t_j, t_{j+k}) = \theta_{j(j+k)}(x_{j(j+k)})$, where $x_{j(j+k)} = t_{j+k} - t_j$, $j = 1, 2 \ldots d-k$ and $k = 1, \ldots d-1$. Then, the copula density in

Equation (2) becomes

$$c(u_1,\ldots,u_d \mid t_1,\ldots,t_d) = \prod_{k=1}^{d-1}\prod_{j=1}^{d-k} c_{j(j+k);S_{jk}}\left(u_{j|S_{jk}}, u_{j+k|S_{jk}}; \theta(x_{j(j+k)})\right). \tag{4}$$

The form of $\theta(\cdot)$ is usually unknown, but can be estimated parametrically or nonparametrically. Here, we take a parametric approach and consider a parametrization in terms of Pearson's correlation coefficient ρ due to its ease of interpretation. Specifically, we propose the following two-parameter exponential decay model

$$\rho_{j(j+k)}\left(x_{j(j+k)};\boldsymbol{\beta}_{j(j+k)}\right) = \exp\left(-\beta_{0,j(j+k)} - \beta_{1,j(j+k)}x_{j(j+k)}\right), \tag{5}$$

where $j = 1,2,\ldots d-k; k = 1,\ldots d-1$, and $\boldsymbol{\beta}_{j(j+k)} = (\beta_{0,j(j+k)}, \beta_{1,j(j+k)})^\top$ is a vector of parameters with $\beta_{0,j(j+k)} > 0$ and $\beta_{1,j(j+k)} > 0$. Under this model, the dependence between two measurements on the same subject decays towards zero at the rate determined by $\beta_{1,j(j+k)}$ as the time interval between the measurements increases, and the level of maximal correlation between two measurements within an infinitesimal time interval is given by $\beta_{0,j(j+k)}$.

Following Killiches and Czado (2018), we can easily transform ρ to Kendall's τ using the relationship $\tau = (2/\pi)\arcsin(\rho)$. Hence, for a specified copula family, the copula parameter θ can be obtained using the one-to-one mapping between θ and τ. Web Table 1 in Web Appendix A under the Supporting Information reports some commonly used parametric copula families along with the Kendall's τ in terms of θ.

3.2 *Data structure*

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}^\top$ be a repeated measurement dataset where $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,d_i})^\top$ is the measurement vector of dimension d_i for subject i, $i = 1,2,\ldots,n$, and let d be the maximum number of measurements in the dataset. Define $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}^\top$ where $\mathbf{t}_i = (t_{i,1}, t_{i,1}, \ldots, t_{i,d_i})^\top$ is the vector of observation times for subject i. Since subjects have different number of measurements in unbalanced longitudinal data, we consider subsets of groups of subjects having the same number of measurements. Let $\boldsymbol{\mathcal{Y}}^g = \{\mathbf{y}_i\,;\, i \in \mathbf{I}_g\}$ denote the collection of observations from n_g subjects having exactly g measurements, where

$\mathbf{I}_g = \{i\,;\, \mathbf{y}_i \in \mathbb{R}^g\}$ is the index set, $g = 1, \ldots, d$. Note that balanced longitudinal data is a special case where all subjects have d measurements. Web Table 2 in Web Appendix A under the Supporting Information illustrates the groupings in a dataset of size $n = 12$ with maximum $d = 5$ measurements per subject. There are $n_1 = 2$, $n_2 = 3, n_3 = 2, n_4 = 2$ and $n_5 = 3$ subjects with $1, 2, 3, 4$ and 5 measurements, respectively. Accordingly, we have $I_1 = \{1, 2\}$, $I_2 = \{3, 4, 5\}$, $I_3 = \{6, 7\}$, $I_4 = \{8, 9\}$, and $I_5 = \{10, 11, 12\}$.

3.3 *Estimation*

For each collection $\boldsymbol{\mathcal{Y}}^g = \{\mathbf{y}_i\,;\, i \in \mathbf{I}_g\}$, the joint distribution of the repeated measurements $\mathbf{y}_i$ is given by Equation (1) where d is replaced by the corresponding dimension g. Let $F_j \equiv F_j(\cdot; \boldsymbol{\alpha}_j \mid t_j)$ be the marginal distribution function of Y_j conditional on t_j with marginal parameter vector $\boldsymbol{\alpha}$, and let $\boldsymbol{\beta}$ denote the parameter vector associated with the D-vine copula $C(\cdot; \boldsymbol{\beta} \mid t_1, \ldots, t_g)$. For notational simplicity, we suppress the dependence on the visit times. Following Equations (1) and (2), the joint density can be written as

$$f(y_1, \ldots, y_g \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{l=1}^{g} f_l(y_l; \boldsymbol{\alpha}_l) \prod_{k=1}^{g-1} \prod_{j=1}^{g-k} c_{j(j+k);S_{jk}} \Big(F_{j|S_{jk}}(y_j|\mathbf{y}_{S_{jk}}; \boldsymbol{\alpha}_{j:j+k-1}, \boldsymbol{\beta}_{j,j+k-1}), F_{j+k|S_{jk}}(y_{j+k}|\mathbf{y}_{S_{jk}}; \boldsymbol{\alpha}_{j+1:j+k}, \boldsymbol{\beta}_{j+1,j+k}); \boldsymbol{\beta}_{j(j+k)} \Big), \tag{6}$$

where $\mathbf{y}_{S_{jk}} = (y_{j+1}, \ldots, y_{j+k-1})$. Then, considering the D-vine construction in Equation (2) for each $g-$dimensional copula, we obtain the log-likelihood function for the full model as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{x}) &= \sum_{l=1}^{d} \sum_{i \in I_g} \log\left(f_l(y_{i,l}; \boldsymbol{\alpha}_l)\right) \\ &\quad + \sum_{g=2}^{d} \sum_{i \in I_g} \sum_{k=1}^{g-1} \sum_{j=1}^{g-k} \log c_{j(j+k);S_{jk}} \Big(F_{j|S_{jk}}(y_{i,j}|\mathbf{y}_{i,S_{jk}}; \boldsymbol{\alpha}_{\{j\} \cup S_{jk}}, \boldsymbol{\beta}_{j,j+k-1}), \\ &\qquad F_{j+k|S_{jk}}(y_{i,j+k}|\mathbf{y}_{i,S_{jk}}; \boldsymbol{\alpha}_{\{j+k\} \cup S_{jk}}, \boldsymbol{\beta}_{j+1,j+k}); \boldsymbol{\beta}_{j(j+k)} \Big) \\ &= \mathcal{L}^M(\boldsymbol{\alpha}|\mathbf{Y}, \boldsymbol{x}) + \mathcal{L}^C(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{x}). \end{aligned} \tag{7}$$

The statistical inference of the above model involves estimation of the marginal part ($\mathcal{L}^M$) and estimation of the copula part ($\mathcal{L}^C$), which will be discussed in the next subsections.

Few aspects are noteworthy. The log-likelihood function in Equation (7) is built using clusters of observations of different size, whereas the special case of balanced longitudinal data consists of one cluster of size n all with d measurements. We assume that the D-vine describing the dependence structure of group g is a sub-vine of the D-vine describing the dependence structure of group $k = g + 1, \ldots, d$. This assumption is reasonable under monotone missingness but would not be warranted under other missing data patterns (see Hasler et al. (2018) for a discussion). While subjects with only one measurement do not contribute to the dependence structure, they are included in the marginal models.

3.3.1 *Estimation of marginal distributions.* We first discuss the univariate marginal modeling of Y_j for $j = 1, \ldots, d$. In general, one can use parametric and nonparametric methods to estimate F_j. Here, we consider the parametric LMM framework which is a popular choice for modeling longitudinal data (Diggle, 2002). However, one can also use other approaches such as the linear model (see Web Appendix B under the Supporting Information for a discussion). Suppose $y_{i,j}$ is the j^{th} measurement on the i^{th} subject, where additional covariate information is available. Let $\mathbf{z}_{i,j} \in \mathbb{R}^p$ and $\mathbf{w}_{i,j} \in \mathbb{R}^q$ be the covariates associated with the fixed and random effects, respectively. The model for $y_{i,j}$ can then be represented as

$$y_{i,j} = \mathbf{z}_{i,j}^\top \boldsymbol{\eta} + \mathbf{w}_{i,j}^\top \boldsymbol{\gamma}_i + \epsilon_{i,j},$$

where $\boldsymbol{\eta} \in \mathbb{R}^p$ is the vector of coefficients for the fixed effects, and $\boldsymbol{\gamma}_i \in \mathbb{R}^q$ is the vector of random effects which is assumed to have a normal distribution with mean vector zero and covariance matrix $\mathbf{B} \in \mathbb{R}^{q \times q}$. Moreover, the error vector $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,d_i})^\top$ for the the i^{th} subject is assumed to be independent of $\boldsymbol{\gamma}_i$ and have a normal distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d_i \times d_i}$. Hence, $y_{i,j} \sim N_d(\mathbf{z}_{i,j}^\top \boldsymbol{\eta}, \phi_{i,j}^2)$ with $\phi_{i,j}^2 = \mathbf{w}_{i,j}^\top \mathbf{B} \mathbf{w}_{i,j} + \sigma_{i,j}^2$ for $j = 1, \ldots, d_i$ and $i = 1, \ldots, n$, where $\sigma_{i,j}^2 = \text{Var}(\epsilon_{i,j})$. The coefficient

vectors $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}_i$, and the covariance matrices $\mathbf{B}$ and $\boldsymbol{\Sigma}_i$ are usually estimated using maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML) (see e.g., Diggle, 2002). Based on the estimates, the pseudo-copula data can be obtained using $\widehat{u}_{i,j} = \widehat{F}_j(y_{i,j}) = \Phi\left(\frac{y_{i,j} - \mathbf{z}_{i,j}^\top \widehat{\boldsymbol{\eta}}}{\phi_{i,j}}\right)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d_i$, where Φ is the cumulative distribution function of the standard normal distribution. The pseudo-copula data $\widehat{\mathbf{U}} = \{\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_n\}^\top$ are then partitioned into groups $\widehat{\boldsymbol{\mathcal{U}}}^g = \{\widehat{\mathbf{u}}_i; i \in \mathbf{I}_g\}$ for $g = 1, 2, \ldots, d$.

3.3.2 *Estimation of dependence parameters.* Given the pseudo-copula data $\widehat{\mathbf{U}}$, the dependence parameters $\boldsymbol{\beta}$ are estimated by maximizing the copula log-likelihood $\mathcal{L}_C(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{x})$ in Equation (7), where $\widehat{\boldsymbol{\alpha}}$ is the estimated marginal parameter vector. While one can use the inference functions for margins (IFM) method (Joe, 2005) to jointly estimate the $d(d-1)$ dependence parameters, this approach becomes computationally infeasible when d gets large. Therefore, the estimation of dependence parameters is typically performed using the sequential approach in Aas et al. (2009). In this approach, one first fits the unconditional pair copulas in tree 1 by maximizing the corresponding pair copula log-likelihoods using the available data. The fitted pair copulas are then used in Equation (3) to obtain the pseudo observations for fitting the pair copulas in tree 2. Using these pseudo observations, the conditional pair copulas in tree 2 are fitted by maximizing the corresponding conditional pair copula log-likelihoods. One continues this way until all dependence parameters are estimated (Aas et al., 2009; Hobæk Haff, 2013). The Akaike information criterion (AIC) (Akaike, 1974) is commonly used to select a copula family for each pair copula (Dißmann et al., 2013).

While these procedures are typically employed for complete data settings, they can be directly adopted to unbalanced longitudinal data assuming responses are missing completely at random (Hasler et al., 2018). We provide the asymptotic results for the IFM and sequential estimators (SEQ) for the time-heterogeneous D-vine copula model focusing on the balanced case in Web Appendix A of the Supporting Information. In the unbalanced case,

the asymptotic behaviour is expected to be the same with a difference that the estimation uncertainty would be propagated for pair copula components having fewer observations.

A schematic representation of the copula $c_{1:5}$ with its pair copulas is illustrated in Figure 2. The repeated measurements are represented by nodes. The associated pair copulas and the available observations for estimation are presented above and below each edge, respectively. Different color intensities of the nodes and different line types of the edges are used to highlight the sub-vines $c_{1:2}$, $c_{1:3}$, $c_{1:4}$ and $c_{1:5}$, where a lighter color indicates more uncertainty. For instance, it can be easily seen that the estimation of c_{12} depends on the observations from not only $\widehat{\boldsymbol{u}}^2$ but also $\widehat{\boldsymbol{u}}^3$, $\widehat{\boldsymbol{u}}^4$ and $\widehat{\boldsymbol{u}}^5$. An illustration of the fitting of the proposed D-vine copula model with at most $d = 5$ repeated measurements can be found in Web Appendix A under the Supporting Information.

[Figure 2 about here.]

3.4 *Parsimonious model specifications*

While the proposed D-vine model in Equation (6) allows pair copulas to have different intercept and decay parameters, one can consider more parsimonious models assuming some of these parameters to be shared at the tree or vine level. Here, we consider two such models. The first uses the same two-parameter exponential decay form for all pair copulas in the vine structure, i.e., $\beta_{0,j(j+k)} = \beta_0$ and $\beta_{1,j(j+k)} = \beta_1$. The second one assumes that each tree has its own two-parameter exponential decay form, i.e., $\beta_{0,j(j+k)} = \beta_{0,k}$ and $\beta_{1,j(j+k)} = \beta_{1,k}$, $j = 1, \ldots, d-k$; $k = 1, \ldots, d-1$. These considerations would be appealing in applications, especially for large d, as fewer parameters need to be estimated. The dependence parameter estimation method can also be tailored under these parsimonious models. For instance, when all pair copulas share the same parameters the estimation needs to be done jointly for the full vine structure. This would correspond to the IFM method (Joe, 2005). On the other hand, when the parameters are shared at the tree level, the parameter estimation needs

to be performed jointly for pair copulas in a given tree and sequentially across trees. In the general case where no such assumption is made, parameters are estimated separately for each pair copula in a sequential manner. Throughout the paper, we refer to the latter approach as the time-heterogeneous D-vine copula model fitted by pair and abbreviate it as HET–P. Similarly, the more parsimonious D-vine copula models fitted sequentially by each tree and fitted jointly for the full vine structure are abbreviated as HET–T and HET–V, respectively.

3.5 *Quantile Prediction*

Consider the measurement vector $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,d_i})^\top$ for subject i. The proposed D-vine copula model can be used to predict the quantile of the $(d_i + 1)$th measurement provided that $d_i + 1 \leqslant d$. For this, we focus on the conditional distribution function $F_{d_i+1|1,\ldots,d_i}(\cdot \mid y_{i,1}, y_{i,2}, \ldots, y_{i,d_i})$. In particular, the conditional quantile function for $p \in (0,1)$ can be expressed as

$$\begin{aligned} q_p(y_{i,1}, y_{i,2}, \ldots, y_{i,d_i}) &= \left(F_{d_i+1|1,\ldots,d_i}\right)^{-1}(p \mid y_{i,1}, y_{i,2}, \ldots, y_{i,d_i}) \\ &= \left(F_{d_i+1}\right)^{-1}\left(C^{-1}_{d_i+1|1,\ldots,d_i}(p \mid u_{i,1}, u_{i,2}, \ldots, u_{i,d_i})\right). \end{aligned} \tag{8}$$

Hence, the conditional quantile function can be written in terms of the inverse marginal distribution function $\left(F_{d_i+1}\right)^{-1}$ of y_{i,d_i+1} and the conditional copula quantile function $C^{-1}_{d_i+1|1,\ldots,d_i}$. For example, in the case of a 5-dimensional D-vine copula, the conditional quantile function $C^{-1}_{5|1,2,3,4}(p|u_1, u_2, u_3, u_4)$ can be expressed as

$$h^{-1}_{5|4}\Bigg[h^{-1}_{5|3;4}\Bigg\{h^{-1}_{5|2;3,4}\Bigg(h^{-1}_{5|1;2,3,4}\Bigg(p \;\Big|\; h_{1|4;2,3}\Big(h_{1|2;3}\big(h_{1|2}(u_1|u_2) \mid h_{3|2}(u_3|u_2)\big) \;\Big|\; h_{4|2;3}\big(h_{4|3}(u_4|u_3) \mid h_{2|3}(u_2|u_3)\big)\Big)\Bigg) \;\Big|\; h_{2|4;3}\big(h_{2|3}(u_2|u_3) \mid h_{4|3}(u_4|u_3)\big)\Bigg) \;\Big|\; h_{3|4}(u_3|u_4)\Bigg\} \;\Big|\; u_4\Bigg],$$

where the h-functions are obtained using Equation (3). Further details on D-vine based quantile predictions can be found in Kraus and Czado (2017).

4. Simulation Study

We conduct extensive simulations to evaluate the performance of the proposed time-heterogeneous D-vine copula model in comparison to those of the time-homogeneous D-vine copula model of Killiches and Czado (2018) and the LMM.

4.1 *Simulation setup*

We consider six data generating processes (DGPs) using D-vine copulas with up to $d = 5$ measurements per subject. In this set of simulations, we focus on the comparisons with the time-homogeneous D-vine copula model. Table 1 summarizes the D-vine copula models used to generate data that consist of bivariate copula families at each tree along with their approximate range of dependence parameters reported in Kendal's τ scale for each DGP.

[Table 1 about here.]

The first three DGPs are constructed under the time-homogeneous (HOM) scenario. In DGP1, we consider each pair copula to have a different constant Kendall's τ value, whereas in DGP2 and DGP3, Kendall's τ is set constant at the tree and vine levels, respectively. The latter three DGPs are constructed under the time-heterogeneous (HET) scenario in a similar fashion. In DGP4, we set a different (β_0, β_1) value for each pair copula of the D-vine, whereas in DGP5 and DGP6, we allow the parameter values to be shared for each tree and for the full vine structure, respectively. For instance, we set $\beta_0 = 0.05$ and $\beta_1 = 2.3$ for all pair copulas in DGP6. The last column of Table 1 lists the expected best fitting model, which corresponds to the underlying DGP.

Under each DGP, we generate $M = 1000$ samples of size $n = 250$ and 500 with $d = 5$ repeated measurements for each subject (balanced setting) and delete the measurements randomly to obtain an unbalanced setting. For each subject i within each sample, we generate visit times t_{ij} from the uniform distribution $U[(j-1)/d, j/d]$ for $j = 1, 2, \ldots, d$. For DGPs under HET scenario, we calculate Pearson's correlation coefficient ρ using Equation (5)

and convert it to first Kendall's τ and then the copula parameter θ under the specified pair copula families. We then generate copula data from the corresponding D-vine copula model for each subject. To mimic the nature of repeated measurements under monotone missingness, we independently draw d_i from a discrete distribution on $(2, \ldots, d)$ for each subject i and restrict this subject to its first d_i components. The proportion of subjects with at least j measurements is roughly 100%, 80%, 60%, 40% for $j = 2, 3, 4, 5$, respectively.

Our aim is to compare different fitting models under each DGP. Specifically, we consider the proposed time-heterogeneous D-vine copula models allowing pair copulas to have different intercept and decay parameters (HET–P), as well as assuming these parameters to be shared for pair copulas at each tree (HET–T) and for the full vine structure (HET–V). We also fit time-homogeneous D-vine copula models assuming a constant dependence parameter for each pair copula (HOM–P), for pair copulas at each tree (HOM–T) and for the full vine structure (HOM–V).

To assess the effect of ignoring time interval on dependencies, we compare Kendall's τ estimates from the fitted models by considering each pair copula separately. For this purpose, we use the mean absolute difference (MAD) between Kendall's τ values of the true model and fitted models for each pair copula calculated as $\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}|\widehat{\tau}^{i}_{fitted}(m) - \tau^{i}_{true}|$. We also compare all fitted models in terms of AIC to facilitate model selection in practical settings.

4.2 *Simulation results: Estimation of dependence parameters*

We calculate the MAD of Kendall's τ averaged over $M = 1000$ simulated balanced and unbalanced datasets of size $n = 250$ and 500 for each of the 10 pair copulas under the six DGPs. The results are presented in Web Tables 4–7 in Web Appendix B of the Supporting Information.

From the results in Web Table 5, we observe that, under DGP1, HOM–P provides the smallest MAD of Kendall's τ for most of the pair copulas with few exceptions where the

smallest MAD of Kendall's τ is achieved by either HOM–T or HOM– V. The MAD values from HET–P are usually very close to the ones obtained from HOM–P. Under DGP2, HOM–T gives the smallest MAD of Kendall's τ for all pair copulas. The second smallest MAD values are usually given by its heterogeneous analogue, HET–T. A similar conclusion is drawn under DGP3 where HOM–V performs the best in terms of the MAD of Kendall's τ for all pair copulas, which is followed by HET–V. The results in Web Table 7 indicate that, under DGP4, the proposed HET–P model yields a smaller MAD of Kendall's τ for most pair copulas than the homogeneous models. In few cases, a smaller MAD is obtained by other time-heterogeneous models. Similarly, for almost all pair copulas under DGP5 and DGP6, the smallest MAD of Kendall's τ is given by HET–T and HET–V, respectively, and this holds for both $n = 250$ and 500.

Overall, the simulation results suggest that the MADs of Kendall's τ are usually quite small for the proposed time-heterogeneous D-vine copula models compared to the time-homogeneous counterparts when there is time heterogeneity in the dependence structure. As expected, the time-homogeneous D-vine copula models tend to achieve the smallest MAD of Kendall's τ when the dependence structure has time homogeneity. However, the proposed time-heterogeneous D-vine copula models yield comparable MAD values to those from the corresponding time-homogeneous D-vine copula models even when the underlying dependence structure is time-homogeneous. Similar conclusions are drawn in the balanced settings (see Web Tables 4 and 6 in Web Appendix B of the Supporting Information).

4.3 *Simulation results: Model selection*

An important aspect in practical settings is to decide among the time-homogenous and time-heterogeneous D-vine copula models that would best fit the data. For this purpose, we consider the AIC as a widely used tool for copula model selection. Web Table 3 in Web Appendix B of the Supporting Information gives the proportion of times out of $M = 1000$

replications that each candidate model has been selected by the AIC under each DGP for sample sizes $n = 250$ and 500. When data are generated under the HOM design, time-homogeneous D-vine copula models fitted by pair, by tree and by vine are most suitable for DGP1, DGP2 and DGP3, respectively. While the model selection accuracy of the AIC is over 95% for HOM–P and HOM–T, it drops to approximately 85% for HOM–V. Similarly, for models under the HET design, time-homogeneous D-vine copula models fitted by pair, by tree and by vine are selected over 95% of the time for DGP4, DGP5 and DGP6, respectively. These results confirm that the AIC is a suitable tool for selecting among time-homogenous and time-heterogeneous D-vine copula models in practice.

4.4 *Additional simulation results: Improvement over LMM*

We further evaluate the performance of the proposed D-vine copula model with that of the LMM focusing on two data generation scenarios. In the first case, data are generated from the LMM with AR(1) correlation structure and auto-correlation parameter $\rho = 0.4$ (DGP0). In the second case, we consider the D-vine copula model under DGP4 (see Table 1). In both cases, we specify the response variable using a binary covariate (Treatment) and a continuous covariate (Time) as:

$$y_{i,j} = \eta_0 + \eta_1 \text{Treatment}_i + \eta_2 \text{Time}_{i,j} + \gamma_i + \epsilon_{i,j},$$

where $\gamma_i \sim N(0, 6.5^2)$, $\epsilon_{i,j} \sim N(0, 3.5^2)$ with $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d_i \leqslant 5$, and $\eta_0 = 4$, $\eta_1 = 3$, and $\eta_2 = 1$. The subject-level covariate Treatment_i is generated from Bernoulli distribution with $p = 0.5$ and $\text{Time}_{i,j}$ is generated from Uniform distribution $U[(j-1)/d, j/d]$ for $j = 1, 2, \ldots, d$. Under each model, we generate $M = 1000$ samples of size $n = 250$ and 500 and compare LMM, HOM–P and HET–P models in terms of their quantile prediction, average AIC and proportion of model selection based on AIC. For quantile prediction, we randomly select a subject with $d = 5$ measurements from each sample and use the corresponding fitted model to obtain the median (50% quantile) and

the 90% prediction intervals (PIs) (5% and 95% quantiles) of the fifth measurement given the first four measurements. We then check whether the true value of the fifth measurement is inside the interval or not. The quantile predictions under the fitted LMM are based on the conditional distribution of the fifth measurement given the previous four measurements, which is univariate Gaussian and its mean and variance are obtained from the fitted LMM using the conditional mean and variance expressions from the standard multivariate Gaussian theory. In Table 2, we report the coverage probability (CP) and the average length of 90% PIs over $M = 1000$ samples of size $n = 250$ and 500 as well as the average AIC values and model selection accuracy based on AIC.

[Table 2 about here.]

Based on the results, the LMM tends to perform better in terms of the average AIC and model selection accuracy under DGP0. However, it fails to achieve the nominal CP and gives wider PIs than HOM–P and HET–P. The proposed time-heterogeneous D-vine copula model yields comparable results to the LMM even when the data are generated under DGP0. Under DGP4, HET–P provides, on average, narrower PIs, smaller AIC values and a higher model selection accuracy for both $n = 250$ and 500. Here, both HOM–P and HET–P models achieve the nominal CP while the LMM suffers from undercoverage. Overall, the simulation results suggest that the proposed time-heterogeneous D-vine copula model achieves better predictive performance compared to its time-homogeneous counterpart as well as the LMM.

5. Data Application: Manitoba Follow-up Study (MFUS)

In this section, we revisit our motivating data example presented in Section 2 and analyze a sub-sample of the MFUS data, which contains 462 participants having at most 5 measurements. Web Table 9 in Web Appendix C under the Supporting Information summarizes the sizes of the group of subjects with exactly j and more than j measurements for $j = 1, \ldots, 5$.

As mentioned in Section 2, we consider the DBP as the response variable for which the covariates of interest are age of the subject (Age), BMI and IHD status. Our main objective is to investigate the effect of time interval on the dependence structure among measurements.

5.1 *Estimation of marginal distribution*

It is clear from Web Figure 3 (see Web Appendix C of the Supporting Information) that the response variable exhibits a relationship with covariates for different visit times such as baseline, visit 1, visit 2 and so on. The MFUS data set has been recently analyzed in Hoque and Torabi (2018) using a mixed model approach. Hence, we employ the LMM to fit the data using the function **lme** from the R library **nlme** (Pinheiro et al., 2013). Specifically, we consider the model

$$y_{i,j} = \eta_0 + \eta_1 \text{Age}_{i,j} + \eta_2 \text{BMI}_{i,j} + \eta_3 \text{IHD}_{i,j} + \gamma_i + \epsilon_{i,j}, \tag{9}$$

where $y_{i,j}$ represents the DBP of subject i at time j, $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,j})^\top \sim N(0, \boldsymbol{\Sigma}_i)$ with $i = 1, 2, \ldots, 462$, $j = 1, 2, \ldots, d_i \leqslant 5$. Here, we consider Age as the time variable having a fixed effect on the response.

A subject-specific random intercept model has been fitted to the data and the following different correlation structures for the error have been compared: (i) general (unstructured): $\text{Cov}(\epsilon_{i,j}, \epsilon_{i,k}) = \sigma_{jk}$; (ii) compound symmetry: $\text{Cov}(\epsilon_{i,j}, \epsilon_{i,k}) = \sigma_\epsilon^2 \rho^{\mathbb{1}\{j \neq k\}}$ for some $\rho \in (-1, 1)$; (iii) AR(1): $\text{Cov}(\epsilon_{i,j}, \epsilon_{i,k}) = \sigma_\epsilon^2 \rho^{|j-k|}$ for some $\rho \in (-1, 1)$; (iv) exponential decay: $\text{Cov}(\epsilon_{i,j}, \epsilon_{i,k}) = \sigma_\epsilon^2 \exp(-|j-k|/d_\epsilon)$, where $d_\epsilon > 0$ denotes the constant range parameter.

We select the best model in terms of the log-likelihood and AIC values (see Table 3), though the parameter estimates are very similar (not shown here). We observe that the best model is the one with exponential decay error structure, which contains intercept, Age, BMI and IHD as fixed effects. The associated model parameter estimates, their standard errors, and their corresponding 95% confidence intervals (CIs) are provided in Web Table 10 in Web Appendix C of the Supporting Information. We observe that Age, BMI and IHD are

positively associated with DBP. The 95% CIs for the fixed effects indicate a significant effect of Age and BMI but not IHD. Nevertheless, we decided to keep IHD in the model as it is clinically important (Rabkin et al., 1978).

5.2 *Estimation of dependence structure*

We transform the data to uniform scale using the fitted LMM with the exponential decay error correlation structure. For this, we apply the univariate conditional normal distribution with the mean and variance estimated from the fitted LMM. The resulting transformed data are then fitted using a D-vine copula model with order 1-2-3-4-5 according to the sequential approach presented in Section 3. Specifically, we fit both the proposed time-heterogeneous D-vine copula models (HET–P, HET–T and HET–V) and time-homogeneous D-vine copula models (HOM–P, HOM–T and HOM–V). In model fitting, we consider parametric pair copulas, and select a copula family for each pair copula from the Gaussian, Frank, Clayton and Gumbel copulas as well as the rotated versions of the latter two using the AIC. The selected copula families under each model are provided in Web Table 11 in Web Appendix C of the Supporting Information.

5.3 *Model Evaluation*

We evaluate the fitted full models (time-heterogeneous and time-homogeneous D-vine copula models including margins) as well as the best fitted LMM using the log-likelihood and AIC. The results are reported in Table 3. As can be seen, HET–P performs the best in terms of the AIC among all fitted models. Moreover, all three time-heterogeneous D-vine copula models yield a better fit than their time-homogeneous counterparts. The LMM shows the worst performance in terms of the AIC indicating that the underlying dependence structure may not be well-represented by a multivariate normal distribution. Based on these comparisons, we can conclude that the time-heterogeneous D-vine copula model which allows different

parameter values for each pair copula provides a better representation of the underlying dependence structure of the data by incorporating the time intervals between measurements.

[Table 3 about here.]

5.4 *Quantile Prediction*

We further compare the fitted HET–P, HOM–P and LMM in terms of their quantile predictions. For this, we consider three subjects with $d_i = 5$ having the following DBP measurements: $\mathbf{y}_1 = (64, 70, 70, 84, 80)^\top$, $\mathbf{y}_2 = (78, 75, 75, 90, 90)^\top$ and $\mathbf{y}_3 = (80, 70, 90, 70, 90)^\top$. We also consider their covariate values $\mathbf{z}_i = (\text{Age}_i, \text{BMI}_i, \text{IHD}_i)^\top$: $\mathbf{z}_1 = (31.00, 25.99, 0)^\top$, $\mathbf{z}_2 = (39.00, 24.41, 0)^\top$ and $\mathbf{z}_3 = (32.17, 25.09, 1)^\top$. To illustrate quantile prediction, we pretend that the selected subjects have only the first four measurements. Then for each fitted model, we obtain the median (50% quantile) and the 90% PI (5% and 95% quantiles) of the fifth measurement $y_{i,5}$ based on the first four measurements $(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4})$. We then compare the predicted values to the observed values of the fifth measurement.

The results are presented in Web Table 12 in Web Appendix C of the Supporting Information. For Subject 1, the observed value of the fifth measurement is inside the PIs under all models. However, HET–P gives a shorter PI than the other two models. For Subject 2, the PI of LMM does not contain the observed value of the fifth measurement, while the PIs from HOM–P and HET–P D-vine copula models both contain the observed value. Similarly, HET–P results in a shorter PI than the other two models. For Subject 3, the observed value of the fifth measurement only falls in the PI of HET–P D-vine copula model. Even though they yield shorter PIs, neither LMM nor HOM–P D-vine copula model can accurately predict the observed value. The PIs are symmetric around mean (median) for LMM due to normality of the conditional distribution. However, this is not the case for HOM and HET copula models as in general the 5% and 95% quantiles are not symmetric around the 50% quantile.

It is also of interest to inspect the influence of covariates on the estimated conditional

quantiles in practice. For instance, in Figure 3, we illustrate the influence of BMI on the median, 5% and 95% quantiles. We observe that the estimated quantiles from LMM depend linearly on BMI and the influence is same for all quantiles as we have the same slope for the all light lines. All PIs have the same widths for different subjects since the standard deviation does not depend on the covariate values. On the other hand, the quantiles estimated from HOM–P and HET–P D-vine copula models do not depend linearly on the BMI value. The length of PIs varies among subjects. We further see that the slopes can vary for different quantiles. Especially, under the HET–P D-vine copula model, the slope can be positive and negative for different quantiles for the same subject.

[Figure 3 about here.]

6. Discussion

In this paper, we provided a general and flexible model to account for both unbalanced study design and irregular spacings of repeated measurements in longitudinal data with a continuous outcome. For this purpose, we proposed an intuitive, computationally efficient and easily interpretable parametric time-heterogeneous D-vine copula model with arbitrary margins that allows for time adjustment in the dependence structure. More specifically, we introduced a two-parameter exponential decay model for Pearson's correlation coefficient to capture the strength of dependence between two measurements. This choice was made for the ease of interpretation and connection to the LMM framework. However, one can specify similar parametric forms for other dependence parameters such as Kendall's τ.

The proposed approach leads to more parsimonious D-vine copula models assuming some parameters to be shared at the tree or vine level. This aspect, along with the parametric specification of time heterogeneity, make our approach appealing for high-dimensional problems. Furthermore, prediction for missing outcomes (e.g., due to dropout) can easily be done

by our approach because of the nested nature of D-vine copula models. Extensive simulations demonstrated that the proposed D-vine copula models not only perform very well in terms of the MAD of Kendall's τ but also lead to an improved predictive performance compared to the LMM and the time-homogeneous D-vine copula model. The proposed approach was illustrated using a subset of the MFUS, where the multivariate Gaussian assumption was questionable and time heterogeneity was present in the dependence structure. Hence, our model was able to provide a better fit to the data than its competitors.

The D-vine copula model for unbalanced longitudinal data is constructed under ignorable monotone missingness assuming data are missing completely at random. Extensions to cases where missing data pattern is not monotone or data are missing at random are feasible but remain challenging (see Hasler et al., 2018, for a discussion).

The focus in this paper was on univariate longitudinal data. It is of interest to extend our proposed D-vine copula model to model dependence in multivariate longitudinal data. Another topic of interest is to develop a flexible and intuitive framework to model dependence in spatial-longitudinal data where measurements are repeatedly collected from the same spatial locations over time. These are some of the topics for future study.

Acknowledgements

The authors would like to thank the Co-Editor, the Associate Editor, and two anonymous referees for constructive comments and suggestions. Funding in support of this work was provided by the Natural Sciences and Engineering Research Council of Canada to Elif Acar (RGPIN 435943-2013 and RGPIN 06753-2020) and Mahmoud Torabi (RGPIN 368462-2016).

Data Availability Statement

The data used in this paper to illustrate our findings are not publicly available but can be obtained by contacting the Director of the MFUS, Dr. Robert Tate at mfus@umanitoba.ca.

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716–723.

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

Diggle, P. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

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.

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

Flebach, N. H., Hebert, P. R., Stampfer, M. J., Colditz, G. A., Willett, W. C., Rosner, B., et al. (1989). A prospective study of high blood pressure and cardiovascular disease in women. *American journal of epidemiology* **130,** 646–654.

Frees, E. W. and Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics* **38,** 360–373.

Hasler, C., Craiu, R. V., and Rivest, L.-P. (2018). Vine copulas for imputation of monotone non-response. *International Statistical Review* **86,** 488–511.

Hobæk Haff, I. (2013). Parameter estimation for pair-copula constructions. *Bernoulli* **19,** 462–491.

Hoque, M. E. and Torabi, M. (2018). Modeling the random effects covariance matrix for longitudinal data with covariates measurement error. *Statistics in Medicine* **37,** 4167–4184.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

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

Jones, R. H. and Ackerson, L. M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77,** 721–731.

Jones, R. H. and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* **47,** 161–175.

Killiches, M. and Czado, C. (2018). A d-vine copula-based model for repeated measurements extending linear mixed models with homogeneous correlation structure. *Biometrics* **74,** 997–1005.

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

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,** 3197–3217.

MacMahon, S., Peto, R., Collins, R., Godwin, J., Cutler, J., Sorlie, P., et al. (1990). Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet* **335,** 765–774.

Mathewson, F. A., Brereton, C. C., Keltie, W. A., and Paul, G. I. (1965). The university of manitoba follow-up study: A prospective investigation of cardiovascular disease: Part i. general description—mortality and incidence of coronary heart disease. *Canadian Medical Association Journal* **92,** 947.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.

Nunez-Anton, V. and Woodworth, G. G. (1994). Analysis of longitudinal data with unequally

spaced observations and time-dependent correlated errors. *Biometrics* **50,** 445–456.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). nlme: Linear and nonlinear mixed effects models. *R package version* **3,** 111.

Rabkin, S. W., Mathewson, A. L., and Tate, R. B. (1978). Predicting risk of ischemic heart disease and cerebrovascular disease from systolic and diastolic blood pressures. *Annals of Internal Medicine* **88,** 342–345.

Rosner, B. and Muǹoz, A. (1988). Autoregressive modelling for the analysis of longitudinal data with unequally spaced examinations. *Statistics in Medicine* **7,** 59–71.

Shi, P., Feng, X., and Boucher, J. P. (2016). Multilevel modeling of insurance claims using copulas. *Annals of Applied Statistics* **10,** 834–863.

Shults, J. and Chaganty, N. R. (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics* **54,** 1622–1630.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. 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,** 1467–1479.

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

Tate, R. B., Cuddy, T. E., and Mathewson, F. A. (2014). Cohort profile: the manitoba follow-up study (mfus). *International Journal of Epidemiology* **44,** 1528–1536.

Whelton, P. K. (1994). Epidemiology of hypertension. *Lancet (London, England)* **344,** 101–106.

World Health Organization (2002). World health report 2002: Reducing risks, promoting healthy life. https://www.who.int/publications/i/item/9241562072.

Supporting Information

The computer code used for the data analysis and simulation examples, and Web Appendices A, B and C of the Supporting Information referenced in Sections 2–5 are available with this paper at the *Biometrics* website on Wiley Online Library.

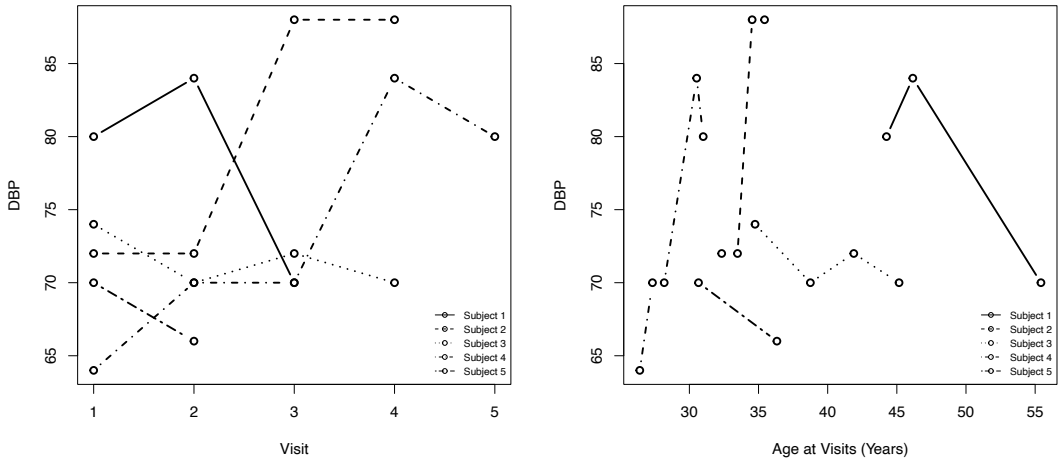

Figure 1. DBP measurements of five subjects from the MFUS across visits (left panel) and age at visits (right panel).

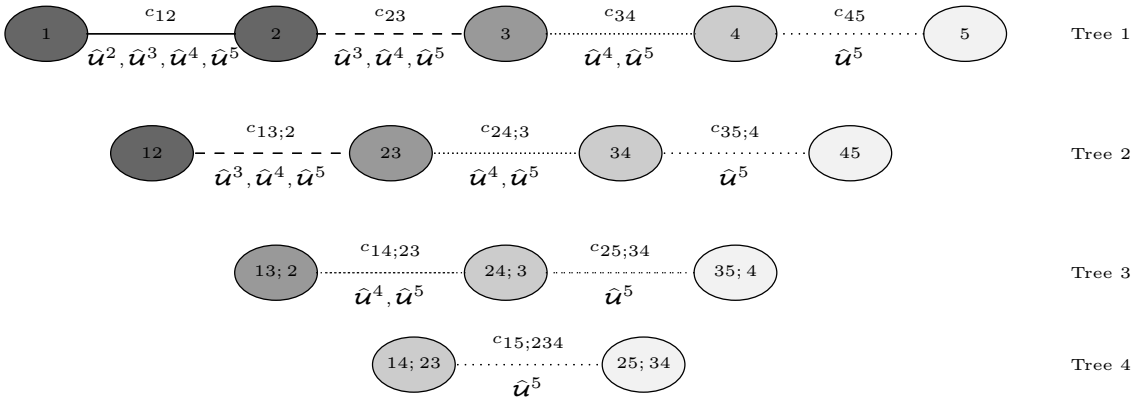

Figure 2. A tree representation of a five-variate D-vine describing the components of the dependence structure of the full model $c_{1:5}$ (dark, medium-dark, medium and light grays). Different colour intensities of nodes and different line types of the edges are used to highlight the sub-vines for $c_{1:2}$ (dark gray nodes; solid line); $c_{1:3}$ (dark and medium-dark gray nodes; solid and dashed lines), $c_{1:4}$ (dark, medium-dark and medium gray nodes; solid, dashed and densely-dotted lines) and $c_{1:5}$ (dark, medium-dark, medium and light gray nodes; solid, dashed, densely-dotted and dotted lines).

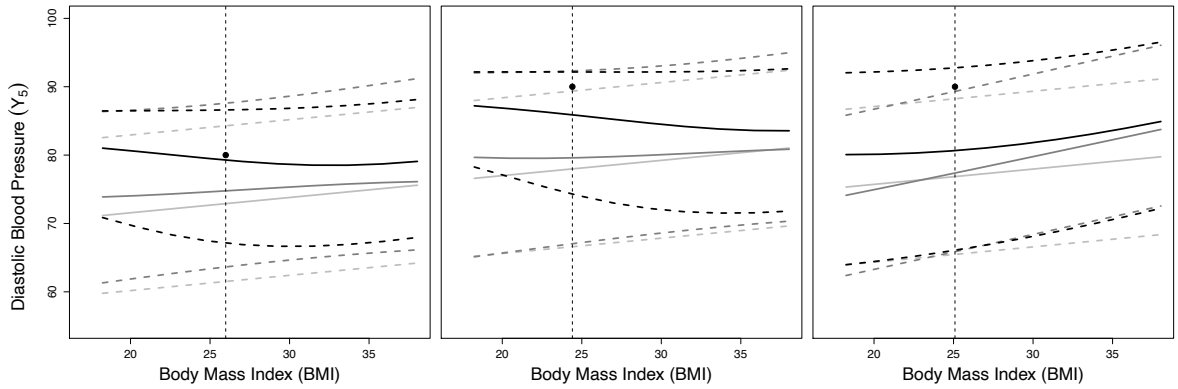

Figure 3. Estimated median (solid lines), 90% prediction intervals (5% and 95% quantiles) (dashed lines) of the fifth measurement for Subject 1 (left panel), Subject 2 (middle panel) and Subject 3 (right panel) for HET–P (dark), HOM–P (medium-dark) and LMM (light). Observed response values are marked as circle. Vertical lines indicate the observed BMI values for the three subjects.

Table 1

D-vine copula models considered in simulations. Listed under each DGP are the bivariate copula families for each tree along with the range of Kendall's τ values (in parenthesis). The first column specifies whether the data generation scenario falls under time-homogeneous (HOM) or time-heterogeneous (HET) setting. The last column lists the expected best fitting model for each DGP.

Scenario	**DGP**	**Tree 1**				**Tree 2**			**Tree 3**		**Tree 4**	**Expected best fitting model**
		c_{12}	c_{23}	c_{34}	c_{45}	$c_{13;2}$	$c_{24;3}$	$c_{35;4}$	$c_{14;23}$	$c_{25;34}$	$c_{15;234}$	
HOM	DGP1	G	C	F	N	G	C	F	N	F	N	HOM–P
		$\tau \in (0.50\ ,\ 0.80)$				$\tau \in (0.30\ ,\ 0.50)$			$\tau \in (0.15\ ,\ 0.30)$		$\tau \in (0\ ,\ 0.15)$	
	DGP2	G	G	G	G	C	C	C	F	F	N	HOM–T
		$\tau = 0.80$				$\tau = 0.50$			$\tau = 0.30$		$\tau = 0.15$	
	DGP3	F	F	F	F	F	F	F	F	F	F	HOM–V
		$\tau = 0.40$				$\tau = 0.40$			$\tau = 0.40$		$\tau = 0.40$	
HET	DGP4	G	C	F	N	G	C	F	N	F	N	HET–P
		$\tau \in (0.50\ ,\ 0.80)$				$\tau \in (0.30\ ,\ 0.50)$			$\tau \in (0.15\ ,\ 0.30)$		$\tau \in (0\ ,\ 0.15)$	
	DGP5	G	G	G	G	C	C	C	F	F	N	HET–T
		$\tau \in (0.50\ ,\ 0.80)$				$\tau \in (0.30\ ,\ 0.50)$			$\tau \in (0.15\ ,\ 0.30)$		$\tau \in (0\ ,\ 0.15)$	
	DGP6	F	F	F	F	F	F	F	F	F	F	HET–V
		$\tau \in (0.25\ ,\ 0.80)$				$\tau \in (0.15\ ,\ 0.40)$			$\tau \in (0.10\ ,\ 0.25)$		$\tau \in (0\ ,\ 0.15)$	

N: Gaussian, C: Clayton, G: Gumbel, F: Frank

Table 2

Coverage probability (CP) and average length of 90% *prediction intervals along with average AIC values and model selection proportions over* $M = 1000$ *samples of size* $n = 250$ *and* 500 *under DGP0 and DGP4.*

		n=250				**n=500**			
	Models	CP	Length	AIC	Selection	CP	Length	AIC	Selection
DGP0	LMM	0.821	11.37	5480.00	0.949	0.858	11.38	10947.18	0.962
	HOM–P	0.900	11.18	5487.74	0.051	0.896	11.28	10955.10	0.038
	HET–P	0.906	11.18	5504.35	0.000	0.898	11.25	10971.39	0.000
DGP4	LMM	0.854	12.75	5723.28	0.000	0.828	12.66	11429.52	0.141
	HOM–P	0.909	11.55	5447.37	0.377	0.906	11.41	11188.92	0.021
	HET–P	0.917	11.42	5443.55	0.623	0.908	11.28	11166.02	0.838

Table 3
Log-likelihood and AIC values with number of parameters for the fitted LMMs with exponential correlation structure and the time-homogeneous and time-heterogeneous D-vine copula models after including margins.

Model	**Log-likelihood**			**AIC**	**# of parameters**	
	Marginal	Copula	Total		Marginal	Dependence
LMM–Exponential	–	–	−5861.2	11736.5	5	2
HOM–P	−5995.3	147.9	−5847.5	11724.9	5	10
HOM–T	−5995.3	137.6	−5857.7	11733.5	5	4
HOM–V	−5995.3	122.2	−5873.1	11758.2	5	1
HET–P	−5995.3	178.8	**−5816.6**	**11683.1**	5	20
HET–T	−5995.3	160.8	−5834.5	11695.0	5	8
HET–V	−5995.3	142.1	−5853.3	11720.6	5	2