

## **Bivariate Copula-based Linear Mixed-effects Models: An Application to Longitudinal Child Growth Data**

P.H. FERREIRA, R.L. FIACCONE, J.S. LORDELO,  
S.O.L. SENA and V.R. DURAN

Received on January 1, 2018 / Accepted on September 25, 2018

**ABSTRACT.** Multiple longitudinal outcomes are common in public health research and adequate methods are required when there is interest in the joint evolution of response variables over time. However, the main drawback of joint modeling procedures is the requirement to specify the joint density of all outcomes and their correlation structure, as well as numerical difficulties in statistical inference, when the dimension of these outcomes increases. To overcome such difficulty, we present two procedures to deal with multivariate longitudinal data. We first present an univariate approach, for which linear mixed-effects models are considered for each response variable separately. Then, a novel copula-based modeling is presented, in order to characterize relationships among the response variables. Both methodologies are applied to a real Brazilian data set on child growth.

**Keywords:** bivariate copula, linear mixed-effects model, longitudinal growth data, time-varying dependence.

### **1 INTRODUCTION**

Longitudinal data or repeated measures data arise when multiple observations are made on the same subject or unit of analysis over time [33].

In practice, it is quite common in clinical trials and social science settings for multiple outcomes to be measured repeatedly within a set of study participants. Some examples are hearing thresholds measured on both ears of a set of participants, HIV studies with CD4 T-cell counts and viral RNA copy numbers are collected longitudinally on each participant, toxicological studies where doses of a toxic agent and some information on its deleterious effect are measured jointly (see [32]). Understanding relationships among multivariate outcomes is challenging due to the complex correlated nature of the problem, whilst providing a unique opportunity in studying the joint evolution of multiple response variables over time.

---

\*Corresponding author: Paulo H. Ferreira – E-mail: paulohenri@ufba.br – <http://orcid.org/0000-0001-6312-6098>

Departamento de Estatística, Universidade Federal da Bahia, Salvador, BA, Brazil.

A number of approaches to joint modeling of multivariate longitudinal data have been proposed in the statistical literature (see, e.g., [4, 6, 12, 17, 24]). When the joint distribution of the two or more outcomes is known, the statistical inference is relatively straightforward using either marginal or joint approaches.

Another possibility is the use of a conditional model, where the joint likelihood of the two or more responses is factorized. The main drawback of these joint modeling procedures is the requirement to specify the joint density of all outcomes or, at least, the correlation structure of the data, which can lead to parsimony and/or computation (optimization) problems, as well as to numerical difficulties in statistical inference, when the dimension of these outcomes increases.

In a multivariate longitudinal study, the relationship between a response and predictors, or the association between a pair of responses, may change over time. In addition to this, the degree of correlation can be of greater interest to the analyst than relation between covariates and marginal mean of outcomes. When the former is the main objective in the multivariate longitudinal analysis, a third alternative for the analysis is to join a set of marginal distributions using a copula function.

This method separates the multivariate joint distribution into two parts: one describing the interdependency of the probabilities, the other describing the marginal distributions alone. Dependence modeling using copula has become very popular the last years (see, e.g., [26, 20]). According to [22], [14], [31] and [8], applications to serial dependence in longitudinal data have a huge potential once the marginal distribution of the process at each point in time can be modelled arbitrarily, while dependence over time is captured by a multivariate copula. This kind of approach is not the primary focus of this paper and will be studied further in future work.

In this paper, we present two procedures, in terms of strengths and weaknesses, to deal with multivariate longitudinal data. In particular, there is an interest in revealing time-varying dependence relationships in the child growth patterns using data from a longitudinal study conducted with 150 children measured monthly, from birth to six months of life. To achieve this goal, we exploit the modularity of copula-based modeling. Considering the applied nature of this work, we make sacrifices to balance the interpretability, complexity and computation of the model. Here, weight and height are outcome variables (i.e. the variables of interest) once were measured monthly, from birth to six months of life.

The paper is organized as follows. In Section 2, we present the univariate and bivariate longitudinal models used in this study. Section 3 concerns statistical inference for them. Section 4 provides an application to a real data set (longitudinal child growth data). Finally, Section 5 ends the paper with some final remarks and directions for future work.

## 2 METHODS

In this section, we present some statistical models that can be used for analyzing multivariate (bivariate, here) longitudinal data. The idea is to show first an univariate approach, where a linear mixed-effects model is considered for each of the two response variables (weight and

height) separately, thus ignoring the (possible) dependence between them. After that, the joint modeling of the two outcomes are based on copulas to relate their marginal distributions at each observation time, and also to describe the dependence structure between them. Therefore, the association between the repeated measurements in a given occasion and the temporal evolution of the anthropometric measures of the individuals are two important aspects that a flexible model should be able to describe.

### 2.1 Univariate Approach

Linear mixed-effects models (also known as multilevel models, hierarchical linear models, random-effects models, among other names), which contain a mixture of fixed effects and random effects, provide a way to deal with longitudinal responses within a subject. The basic idea is to model additional sources of variability in the response by introducing additional terms into the relationship between mean response and explanatory variables, which are random quantities rather than fixed parameters. According to [7], these models consist of two parts: one represents intra-individual (or within-subject) change; and the other represents individual differences in these changes (between-subject or inter-individual variation). The inclusion of subject-specific random effect allows for heterogeneity between subjects and also induces within-subject correlation into the model [32].

Let  $Y_{it}$  be the response of the  $i$ -th child at time  $t$ , with  $i = 1, 2, \dots, m$  and  $t = 0, 1, \dots, n_i$ . In this work,  $m = 150$ ,  $n_i$  varies between two and six, and it is considered two response variables: the height ( $Y_{1it}$ ; in centimeters, cm) and the weight ( $Y_{2it}$ ; in grams, g) of the studied children, which were collected monthly from birth ( $t = 0$ ) to six months of life ( $t = 6$ ).

A special case of the mixed-effects model is the random intercept model, which can be viewed as the deviation of the  $i$ -th subject-specific of the outcome from the population mean of the response. Although the simplicity of the mixed model with only random intercept is appealing, it poses the restriction that the correlation between the repeated measurements remains constant over time. An extension of this model is a mixed model with random intercept and slope, where an additional random-effects term is included, describing that the rate of change in covariates differs between subjects. In general, the linear mixed-effects model can be expressed as follows:

$$Y_{it} = X'_{it}\beta + Z'_{it}\omega_i + \varepsilon_{it}, \tag{2.1}$$

where  $X_{it} = (X_{it}^{(0)}, X_{it}^{(1)}, \dots, X_{it}^{(p-1)})'$  is a known set of  $p$  covariates for the  $i$ -th child, associated with the unknown fixed-effects parameters  $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})'$ . In our case, the time-fixed (or time-invariant) characteristics that make up the  $n_i \times p$  design matrix  $X_i = [X_i^{(0)} \ X_i^{(1)} \ \dots \ X_i^{(p-1)}]$  are: vector of ones ( $X_i^{(0)} = 1$ , due to the fixed intercept), child's gender ( $X_i^{(1)}$ ; female or male) and total duration of breastfeeding (TDB) ( $X_i^{(2)}$ ; in months), in addition to the time-dependent variables: child's age ( $X_i^{(3)}$ ; corresponding month) and child's hemoglobin level ( $X_i^{(4)}$ ; in grams per deciliter, g/dl). Thus,  $p = 5$ , here (four covariate coefficients -  $\beta_1, \beta_2, \beta_3, \beta_4$  - plus the fixed

intercept -  $\beta_0$ ). Moreover,  $Z_{it} = (Z_{it}^{(0)}, Z_{it}^{(1)}, \dots, Z_{it}^{(q-1)})'$  is a set of  $q$  covariates that characterizes random variation in the response attributable to among-unit sources, and it is associated with the random-effects  $\omega_i = (\omega_{i0}, \omega_{i1}, \dots, \omega_{iq-1})'$ , which characterize among-unit variation (specific to child  $i$ ). Finally,  $\varepsilon_{it}$  represents the random error, which describes variation due to sources, like within-unit fluctuations and measurement error. The usual assumptions are:  $\omega_i$  has a multivariate ( $q$ -variate) normal distribution with zero mean vector and covariance matrix  $\Sigma$ , i.e.  $\omega_i \sim N_q(0, \Sigma)$ , and  $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{in_i})'$  has a multivariate ( $n_i$ -variate) normal distribution with zero mean and covariance matrix (diagonal)  $R_i = \sigma^2 V_i$ , i.e.  $\varepsilon_i \sim N_{n_i}(0, R_i)$ . It is also assumed that random effects and random error are both independent across subjects given the covariates. Furthermore, the most common choice for  $R_i$  is the model that says variance is the same at all time points for all units, i.e.  $R_i = \sigma^2 I_{n_i}$ , where  $I_{n_i}$  stands for the identity matrix of order  $n_i$ .

It is important to highlight that additional correlation among the errors can be accommodated by allowing for a more general covariance structure in the model. [7] suggest an additive decomposition of  $\varepsilon_i$  into serially correlated variation and measurement error. It is assumed that  $\varepsilon_i = \varepsilon_{(1)i} + \varepsilon_{(2)i}$ , where  $\varepsilon_{(1)i}$  is a component of serial correlation, which is usually a decreasing function of the time separation, and  $\varepsilon_{(2)i}$  is an extra component of measurement error reflecting variation added by the measurement process itself. According to the same authors, parsimonious choice of the covariance structure can improve the efficiency of inferences made about mean structure and obtain better estimates of standard errors of estimated mean parameters. An additional property of the linear mixed model is related to the covariance of the response profile, which can be described in terms of a set of covariance parameters in both matrix  $\Sigma$  and  $R_i$  [12]. The most commonly used covariance structures are shown below:

- *Compound symmetry structure*, i.e. the  $(k, l)$ -th entry of  $R_i$  is  $\sigma^2 \rho^{1(k \neq l)}$ , for some  $\rho \in [-1, 1]$ , where  $1(\cdot)$  is an indicator function, which in this case takes value 1 when  $k \neq l$  and value 0 when  $k = l$ ;
- *Autoregressive structure of order 1 (AR(1))*, i.e. the  $(k, l)$ -th entry of  $R_i$  is given by  $\sigma^2 \rho^{|k-l|}$ , for some  $\rho \in [-1, 1]$ . This structure is widely used for fitting models to data sets with equally spaced longitudinal observations on the same units of analysis [33];
- *Toeplitz structure*, which specifies that covariance depends only on lag, i.e. the  $(k, l)$ -th entry of  $R_i$  is given by  $\sigma_{kl}^{1(k \neq l)}$  or  $R_i = \sigma^{21(k=l)}$  otherwise;
- *Exponential decay structure*, i.e. the  $(k, l)$ -th entry of  $R_i$  is given by  $\sigma^2 \exp\{-|k-l|/r\}$ , where  $r > 0$  is the constant range parameter.

The number of parameters of the covariance matrix depends on its structure. This generality, however, brings the obvious disadvantage of having a very large number of parameters [23]. In this sense, it is necessary to determine the best structure of the covariance matrix in the data modeling. However, there are no general simple techniques available to compare all these models. [7] suggest to use the empirical variogram of the residuals as a tool for the selection of the

covariance structure, especially when the data are irregularly spaced and unbalanced. According to [12], when data are highly unbalanced with many repeated measurements per subject, a reasonable strategy is to use a simple covariance structure for  $\epsilon_i$ , once random effects can account for most of the variation in the data.

The simplest cases of linear mixed models are random intercept models, where there is a single normally distributed random effect. Random intercept models are sometimes referred to as variance components, error components, or random-effects models. Thus, in our context, a random intercept model for the response variable height can be expressed as follows:

$$Y_{1it} = \beta_{10} + \beta_{11}X_{it}^{(1)} + \beta_{12}X_{it}^{(2)} + \beta_{13}X_{it}^{(3)} + \beta_{14}X_{it}^{(4)} + \omega_{10i} + \epsilon_{1it}, \tag{2.2}$$

where  $Y_{1it}$  is the height of the child  $i$  at time  $t$ ,  $\beta_{10}$  is the fixed intercept,  $\beta_{11}$  is the gender effect,  $\beta_{12}$  is the TDB effect,  $\beta_{13}$  is the age effect,  $\beta_{14}$  is the hemoglobin level effect,  $\omega_{10i} \sim N(0, \sigma_{\omega_{10}}^2)$  is the random intercept and  $\epsilon_{1it}$  is the error.

The mixed-effects model with random intercept and slope, for the response variable weight, can be expressed as:

$$Y_{2it} = \beta_{20} + \beta_{21}X_{it}^{(1)} + \beta_{22}X_{it}^{(2)} + \beta_{23}X_{it}^{(3)} + \beta_{24}X_{it}^{(4)} + \omega_{20i} + \omega_{23i}X_{it}^{(3)} + \epsilon_{2it}, \tag{2.3}$$

where  $Y_{2it}$  is the weight of the child  $i$  at time  $t$ ,  $\beta_{20}$  is the fixed intercept,  $\beta_{21}$  is the gender effect,  $\beta_{22}$  is the TDB effect,  $\beta_{23}$  is the age effect,  $\beta_{24}$  is the hemoglobin level effect,  $\omega_{20i} \sim N(0, \sigma_{\omega_{20}}^2)$  and  $\omega_{23i} \sim N(0, \sigma_{\omega_{23}}^2)$  are the (uncorrelated) random effects of the intercept and slope, respectively, and  $\epsilon_{2it}$  is the error.

In this work, the AR(1) and compound symmetry covariance structures were assumed for the errors  $\epsilon_{1i} = (\epsilon_{1i1}, \epsilon_{1i2}, \dots, \epsilon_{1im_i})'$  and  $\epsilon_{2i} = (\epsilon_{2i1}, \epsilon_{2i2}, \dots, \epsilon_{2im_i})'$  of the linear mixed-effects models (2.2) and (2.3), respectively. These choices were made based on the values of the Akaike Information Criterion (AIC) [1] and through the construction of the semi-variogram [7].

## 2.2 Multivariate Approach

A number of approaches to joint modeling of multivariate longitudinal data have been proposed in the statistical literature (see, e.g., [4, 6, 12, 17, 24]). The analysis can be demanding because of the existence of correlations between multiple time-dependent responses repeated over time.

As an alternative approach, copulas have been widely used when the interest resides in modeling the dependence structure among variables. In practice, copulas have shown to be a good option when the assumption of multivariate normality of the data is doubtful. Moreover, they can capture nonlinear dependence and tail dependence, and have no constraints on the marginal distributions of random variables (great flexibility). Thus, copulas have been successfully applied in the areas of finance, actuarial science and biomedical studies [20], as well as in engineering, in multivariate process control and hydrological modeling [34, 16]. In our work, the copula-based approach allows us to model dependence between the height and weight of 150 children in the study.

In probability theory and statistics, a copula  $C$  is a multivariate distribution function whose marginal distribution functions have their domains belonging to the real range  $[0, 1]$ . In the bivariate context, we have  $C(u, v) = P(U \leq u, V \leq v)$ , where  $U$  and  $V$  are both random variables uniformly distributed over the  $(0, 1)$  interval (i.e.  $U, V \sim U(0, 1)$ ) and can be originated from transformations (probability integral transforms) of any continuous random variables, say  $X$  and  $Y$ . In our paper, we have considered that  $X$  represents the heights of the children, while  $Y$  represents their weights, where  $X$  and  $Y$  have distribution functions  $F(x)$  and  $G(y)$ , respectively.

Copulas can be formally defined as follows.

**Definition 1 (Copula).** *A bivariate copula is a function  $C : [0, 1]^2 \rightarrow [0, 1]$  satisfying the following requirements:*

- *Grounded:*  $C(u, 0) = C(0, v) = 0$ ;
- *Uniform marginals:*  $C(u, 1) = u$  and  $C(1, v) = v$ ;
- *2-increasing:*  $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$  for all  $u_2 > u_1$  and  $v_2 > v_1$ .

Sklar's theorem is the most important result about copulas. The bivariate version of this theorem is presented below.

**Theorem 1.** [30] *Let  $X$  and  $Y$  be two random variables with distribution functions  $F(x)$  and  $G(y)$ , respectively, and joint distribution function  $H(x, y)$ . Then, there exists a copula  $C$  such that, for all  $x, y \in \mathbb{R}$ , satisfies  $H(x, y) = C(F(x), G(y))$ . If  $F$  and  $G$  are continuous, then  $C$  is unique.* The theorem above leads us to conclude that any bivariate distribution function can be constructed through the combination of its marginals by means of a copula [26].

In Table 1, we can see the different copula functions,  $C(u, v)$ , and their densities,  $c(u, v)$ , which are used in this paper, as well as the copula association parameter (and its possible values), denoted by  $\theta$ . For instance, Table 1 shows the bivariate Gaussian copula and Student's  $t_\nu$  copula, as well as their density functions, where  $\Phi^{-1}(\cdot)$  and  $T_\nu^{-1}(\cdot)$  denote the quantile functions of the standard normal and Student's  $t$  distribution with  $\nu$  degrees of freedom, respectively, and  $\Gamma(\cdot)$  is the gamma function.

It is important to mention that the copula captures all the dependence information between the variables. The copula-based dependence measures have been studied for decades. E.g., [29] proved many mathematical properties for some copula-based dependence measures.

Another important definition is given below. It shows us how the copula densities presented in Table 1 (third column) were obtained.

**Definition 2 (Copula density).** *If  $C(u, v) = \int_0^u \int_0^v \frac{\partial^2 C(s, t)}{\partial s \partial t} dt ds$ , for all  $(u, v) \in [0, 1]^2$ , then  $C$  is said to be absolutely continuous and the copula density  $c$  can be defined by  $c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$ .*

Linear correlation (or Pearson's product-moment correlation) coefficient is most frequently used in practice as a measure of dependence. However, in general, it is not possible to construct a

Table 1: Some copula functions, their densities and parameter space.

Copula	$C(u, v)$	$c(u, v)$	$\theta \in$
Gaussian	$\int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{\frac{1}{2}}} \times \exp\left\{-\frac{s^2-2\theta st+t^2}{2(1-\theta^2)}\right\} dt ds$	$\sqrt{\frac{1}{1-\theta^2}} \exp\left\{-\frac{\zeta_1^2-2\zeta_1\zeta_2+\zeta_2^2}{2(1-\theta^2)}\right\} \exp\left\{\frac{1}{2}(\zeta_1^2+\zeta_2^2)\right\}$ where $\zeta_1 = \Phi^{-1}(u)$ and $\zeta_2 = \Phi^{-1}(v)$	$[-1, 1]$
Student's $t_\nu$	$\int_{-\infty}^{T_V^{-1}(u)} \int_{-\infty}^{T_V^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{\frac{1}{2}}} \times \left(1 + \frac{s^2-2\theta st+t^2}{\nu(1-\theta^2)}\right)^{-\frac{\nu+2}{2}} dt ds$	$\frac{1}{2} \left(\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})}\right) \nu(1-\theta^2)^{-\frac{1}{2}} \times \left[\left(1 + \frac{[T_V^{-1}(u)]^2}{\nu}\right) \left(1 + \frac{[T_V^{-1}(v)]^2}{\nu}\right)\right]^{\frac{\nu+1}{2}} \times \left(1 + \frac{[T_V^{-1}(u)]^2-2\theta T_V^{-1}(u)T_V^{-1}(v)+[T_V^{-1}(v)]^2}{(1-\theta^2)\nu}\right)^{-1-\frac{\nu}{2}}$	$[-1, 1]$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$	$(1+\theta)(uv)^{-1-\theta} (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-\theta}$	$(0, \infty)$
Frank	$-\frac{1}{\theta} \log\left(1 + \frac{[\exp\{-\theta u\}-1][\exp\{-\theta v\}-1]}{\exp\{-\theta\}-1}\right)$	$\frac{\theta[1-\exp\{-\theta\}]\exp\{-\theta(u+v)\}}{1-\exp\{-\theta\}-(1-\exp\{-\theta u\})-(1-\exp\{-\theta v\})}$	$\mathbb{R} - \{0\}$
Gumbel	$\exp\left\{-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{\frac{1}{\theta}}\right\}$	$\frac{(-\log u)^{\theta-1}(-\log v)^{\theta-1}}{uv} \times \exp\left\{-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{\frac{1}{\theta}}\right\} \times \left(\left[(-\log u)^\theta + (-\log v)^\theta\right]^{\left(\frac{1-\theta}{\theta}\right)^2}\right) + \left([\theta - 1] \left[(-\log u)^\theta + (-\log v)^\theta\right]^{\left(\frac{1-2\theta}{\theta}\right)}\right)$	$[1, \infty)$

joint distribution of the margins with arbitrary linear correlation coefficient, once this measure does not completely determine the joint distribution. According to [19], the population versions of Kendall's tau ( $\tau_K$ ) and Spearman's rho ( $\rho_S$ ) can be represented in terms of copulas. Thus, we have chosen for this work measures of dependence that are copula-based. To illustrate measures of a form of dependence known as concordance, that are invariant under strictly monotone transformations of the random variables, let us consider the following definition.

**Definition 3 (Concordance).** (i) Two observations  $(x_1, y_1)$  and  $(x_2, y_2)$  are concordant if  $x_1 \leq x_2$  and  $y_1 \leq y_2$  or if  $x_1 \geq x_2$  and  $y_1 \geq y_2$ . An equivalent characterization is  $(x_1 - x_2)(y_1 - y_2) > 0$ . The observations  $(x_1, y_1)$  and  $(x_2, y_2)$  are said to be discordant if  $(x_1 - x_2)(y_1 - y_2) < 0$ . (ii) If  $C_1$  and  $C_2$  are copulas, we say that  $C_1$  is less concordant than  $C_2$  (or  $C_2$  is more concordant than  $C_1$ ) if  $C_1(u, v) \leq C_2(u, v)$  for all  $(u, v) \in [0, 1]^2$ .

Measures of concordance assume the maximum (minimum) value +1 (-1) if the support of the joint distribution function of  $X$  and  $Y$  contains only concordant (discordant) pairs. The measures Kendall's tau and Spearman's rho have been chosen in this work, and they can be expressed in terms of copulas as  $\tau_K = \tau_K(X, Y) = 4 \int \int_{[0,1]^2} C(u, v) dC(u, v) - 1$  and  $\rho_S = \rho_S(X, Y) = 12 \int \int_{[0,1]^2} C(u, v) dudv - 3$ , respectively. Their expressions, corresponding to each copula model chosen here, can be observed in Table 2.

Another important class of dependence measures in nonlinear context is tail dependence. According to [10], the concept of tail dependence relates to the amount of dependence in the right

upper quadrant or left lower quadrant of a bivariate distribution, and it is relevant for the study of dependence between extreme values. The same authors emphasize that the tail dependence between two continuous random variables is a copula property and the amount of tail dependence is invariant under strictly increasing transformations. The definition of tail dependence is shown below.

**Definition 4 (Tail dependence).** Let  $C$  be the copula of  $X$  and  $Y$ .

The lower tail dependence parameter is  $\lambda_L = \lambda_L(X, Y) = \lim_{t \rightarrow 0^+} P(Y \leq G^{-1}(t) | X \leq F^{-1}(t)) = \lim_{t \rightarrow 0^+} \frac{C(t, t)}{t}$ .

The upper tail dependence parameter is  $\lambda_U = \lambda_U(X, Y) = \lim_{t \rightarrow 1^-} P(Y > G^{-1}(t) | X > F^{-1}(t)) = \lim_{t \rightarrow 1^-} \frac{1 - 2t + C(t, t)}{1 - t}$ .

The expressions for the tail dependence parameters (or coefficients), corresponding to each copula model, can be seen in Table 2.

Table 2: Some copula models and their measures of dependence.

Copula	$\tau_K$	$\rho_S$	$\lambda_L$	$\lambda_U$
Gaussian	$\frac{\arcsin(\theta)}{\pi/2}$	$\frac{\arcsin(\theta/2)}{\pi/6}$	0	0
Student's $t_\nu$	$\frac{\arcsin(\theta)}{\pi/2}$	$\frac{\arcsin(\theta/2)}{\pi/6}$	$2T_{\nu+1} \left( -\sqrt{\frac{(\nu+1)(1-\theta)}{1+\theta}} \right)^\ddagger$	$2T_{\nu+1} \left( -\sqrt{\frac{(\nu+1)(1-\theta)}{1+\theta}} \right)$
Clayton	$\frac{\theta}{\theta+2}$	No closed-form	$2^{\frac{1}{\theta}}$	0
Frank	$1 - \frac{4[D_1(\theta)-1]}{\theta} \dagger$	$1 - \frac{12[4D_1(\theta)-2D_2(\theta)]}{\theta}$	0	0
Gumbel	$\frac{\theta-1}{\theta}$	Complicated	0	$2 - 2^{\frac{1}{\theta}}$

$\dagger D_m(\theta) = (m/\theta^m) \int_0^\theta t^m / (e^t - 1) dt$  is the so-called "Debye" function of order  $m$ ,  $m = 1, 2$ .

$\ddagger T_{\nu+1}(\cdot)$  is the cumulative distribution function (cdf) of the Student's  $t$  distribution with  $(\nu + 1)$  degrees of freedom.

One class of multivariate distributions that enable modeling of extremes and other forms of non-normal dependence is the elliptical distributions. Further details on elliptical distributions can be found in [11]. *Elliptical copulas*, which are of great interest here, are simply the copulas of elliptical distributions and provide a rich source of multivariate distributions that share many of the tractable properties of the multivariate normal (or Student's  $t$ ) distribution. Among the elliptical copulas, we highlight the bivariate Gaussian and Student's  $t$  copulas, whose main characteristics are presented in Tables 1 and 2. Notice from Table 2 that the Gaussian copula can not accommodate tail dependence, while the Student's  $t$  copula has lower and upper tail dependence of the same magnitude.

Another important and interesting parametric family of copulas is *Archimedean*. This class allows a great variety of dependence structures. Furthermore, in contrast to elliptical copulas, all commonly encountered Archimedean copulas have closed-form expressions [10]. The main bivariate Archimedean copulas (Clayton copula [5], Frank copula [13] and Gumbel copula [18]) are shown in Table 1. The Kendall's tau and Spearman's rho measures for these Archimedean copulas are presented in Table 2, except for the Clayton and Gumbel copulas, which have no closed-form expressions for the Spearman's rho. Note from Table 2 that the Gumbel copula is able to model

upper tail dependence, whereas the Clayton copula can model lower tail dependence and the Frank copula does not accommodate tail dependence.

The parametric families of copula that were presented here, are important when selecting a good candidate copula model that summarizes the dependence structure between the height and weight of 150 studied children, according to the chosen measures of concordance.

### 3 INFERENCE

In this section, we discuss inference (point and interval estimation) for the parameters of the proposed univariate and bivariate longitudinal data models.

#### 3.1 Point Estimation

The likelihood function for the linear mixed-effects model (2.1) is given by

$$L(\delta|y) = \prod_{i=1}^m \int f(y_i|\omega_i, \beta, R_i, \Sigma) f^*(\omega_i|\Sigma) d\omega_i, \tag{3.1}$$

where  $\delta = (\beta, R, \Sigma)$  is the set of parameters to be estimated, with  $R$  being the  $(\sum_{i=1}^m n_i) \times (\sum_{i=1}^m n_i)$  block diagonal matrix whose elements are  $R_i$ , for  $i = 1, 2, \dots, m$ ;  $y = (y_{11}, \dots, y_{mm})'$  is the  $(\sum_{i=1}^m n_i) \times 1$  vector of all observed responses, with  $y_i = (y_{i1}, \dots, y_{in_i})'$  being the  $n_i \times 1$  vector of observed responses for child  $i$ ;  $f(y_i|\omega_i, \beta, R_i, \Sigma)$  represents the conditional density of response vector  $Y_i$  at point  $y_i$  given  $\omega_i$ ; and  $f^*(\omega_i|\Sigma)$  is the joint density of  $\omega_i$ . It follows from Section 2.1 that  $Y_i|\omega_i \sim N_{n_i}(X_i\beta + Z_i\omega_i, R_i)$  and  $\omega_i \sim N_q(0, \Sigma)$ , where  $Z_i$  is the  $n_i \times q$  matrix defined as  $Z_i = [Z_i^{(0)} \ Z_i^{(1)} \ \dots \ Z_i^{(q-1)}]$ . Notice that (3.1) is the marginal density function of  $y$ , which is obtained by integrating the joint density of  $y$  and  $\omega$  over  $\omega$ , where  $\omega = (\omega'_1, \dots, \omega'_m)'$ .

In order to find the maximum likelihood estimates (MLEs) for the parameters of model (2.1), we set the score function to zero (this function is defined as the first order partial derivative of the logarithm of likelihood function (3.1) with respect to  $\delta$ ). However, it is not always possible to find closed-form expressions for these estimators. Therefore, the use of iterative methods is often needed.

Nevertheless, the usual method of maximum likelihood estimation provides biased estimates for the variance components, in the presence of random effects. An alternative is to use the restricted maximum likelihood (RML) estimation method (sometimes called residual maximum likelihood estimation). According to [7], the RML method is used for unbiased estimation of the variance components in a generalized linear mixed model. This approach consists of dividing the observations into two independent parts: one refers to the fixed effects and the other to the random effects, such that the observed probability density function is obtained by summing those parts. In this work, we perform the RML estimation method for the proposed linear mixed-effects models (2.2) and (2.3), by using the `lme` function from the `n1me` package of R software [27].

The log-likelihood function for the bivariate copula-based linear mixed-effects models, on the other hand, is given by the composition of Eq. (3.1) and the third column of Table 1, as follows:

$$\ell(\eta) = \sum_{i=1}^m \sum_{t=1}^{n_i} \log c(u_{it}, v_{it} | \theta) + \sum_{j=1}^2 \log L_j(\delta_j | y_j), \tag{3.2}$$

where  $\eta = (\delta_1, \delta_2, \theta)$  is the set of all model parameters,  $\delta_j = (\beta_j, R_j, \Sigma_j)$  is the margin  $j$ 's parameter set (here,  $j = 1$  refers to the linear mixed-effects model (2.2), while  $j = 2$  corresponds to the linear mixed-effects model (2.3)),  $y_j = (y_{j11}, \dots, y_{jm_m})$  is the margin  $j$ 's  $(\sum_{i=1}^m n_i) \times 1$  vector of all observed responses,  $u_{it} = F(y_{1it} | \delta_1)$  and  $v_{it} = G(y_{2it} | \delta_2)$  represent the cdf for the linear mixed-effects models (2.2) and (2.3), respectively. Here,  $R_1$  has the AR(1) structure and  $R_2$  exhibits the compound symmetry pattern (see Section 2.1). Moreover,  $L_j(\delta_j | y_j)$  denotes the likelihood function (3.1) for margin  $j$ 's model.

Regarding estimation of parameters in copula-based models, we can use a fully parametric estimation method called Inference Function for Margins (IFM) by [21]. In their work, the authors highlight some advantages of this two-stage maximum likelihood estimation approach: (i) the IFM method is very useful for many multivariate models computationally unfeasible (at first glance); (ii) it allows one to make inference and modeling, starting with univariate and lower-dimensional margins; (iii) there is some robustness against misspecification of the dependence structure, as well as more robustness against outliers or perturbations of the data, compared with the maximum likelihood method; (iv) sparse multivariate data can create problems for the maximum likelihood method, but the IFM method avoids the sparseness problem to a certain degree, especially if all parameters can be estimated from univariate and bivariate likelihoods (this can be a major advantage in a smaller sample situation).

In the first step, the IFM method estimates the marginal parameters  $\delta_j, j = 1, 2$ , through

$$\hat{\delta}_j = \arg \max_{\delta_j} \log L_j(\delta_j | y_j).$$

Then, we obtain  $\hat{u}_{it} = F(y_{1it} | \hat{\delta}_1)$  and  $\hat{v}_{it} = G(y_{2it} | \hat{\delta}_2)$ , in order to estimate the association parameter  $\theta$  by using the pseudo log-likelihood as follows:

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^m \sum_{t=1}^{n_i} \log c(\hat{u}_{it}, \hat{v}_{it} | \theta).$$

Moreover, we consider time-varying bivariate copula-based linear models for fitting the child growth data in the sense that the copula association parameter is not constant, i.e. it varies with time. The reason for this is that the association (or dependence) between the two variables of interest (height and weight) may vary across time, as well as the covariate effects on them and variance components. For further details on time-varying copulas, we refer the reader e.g. to the survey paper by [25].

The log-likelihood function for the time-varying bivariate copula-based linear models can be written as follows:

$$\ell(\xi) = \sum_{i=1}^m \sum_{t=1}^{n_i} \log c(u_{it}, v_{it} | \theta_t) + \sum_{i=1}^m \sum_{t=1}^{n_i} \sum_{j=1}^2 \log f_{jt}(y_{jit} | \delta_{jt}), \tag{3.3}$$

where  $\xi = (\delta'_1, \delta'_2, \theta')$  is the set (vector) of all model parameters;  $\delta_j = (\delta'_{j1}, \dots, \delta'_{jn_i})'$ , with  $\delta_{jt} = (\beta'_{jt}, \sigma_{jt}^2)'$ , is the margin  $j$ 's parameter vector;  $\beta_{jt}$  and  $\sigma_{jt}^2$  are, respectively, the vector of regression coefficients and the variance for margin  $j$  at time point  $t$  (time-varying coefficients and variance);  $\theta = (\theta_1, \dots, \theta_{n_i})'$  is the vector of copula parameters (time-varying association parameter);  $f_{jt}(y_{jit} | \delta_{jt})$  represents the density of response  $Y_{jit}$  at point  $y_{jit}$ ; finally,  $u_{it} = F_t(y_{1it} | \delta_{1t})$  and  $v_{it} = G_t(y_{2it} | \delta_{2t})$  represent, respectively, the cdf for the linear models in margins 1 and 2 at time point  $t$ . Note that, in this case, we fit a linear model to each margin at each time point, i.e. we consider the (classical) multiple linear regression model:  $Y_{jit} = X'_{jit} \beta_{jt} + \varepsilon_{jit}$ , where  $\varepsilon_{jit} \sim N(0, \sigma_{jt}^2)$ , for  $j = 1, 2, i = 1, \dots, m$  and  $t = 1, \dots, n_i$ . This implies that  $Y_{jit} \sim N(X'_{jit} \beta_{jt}, \sigma_{jt}^2)$ .

For model estimation, the log-likelihood function form given by (3.3) also enables the use of the IFM method, which estimates the marginal parameters  $\delta_{jt}$ , for  $j = 1, 2$  and  $t = 1, \dots, n_i$ , at a first step through

$$\hat{\delta}_{jt} = \arg \max_{\delta_{jt}} \sum_{i=1}^m \log f_{jt}(y_{jit} | \delta_{jt})$$

and then estimates the association parameters  $\theta_t, t = 1, \dots, n_i$ , given  $\hat{\delta}_{jt}$  by

$$\hat{\theta}_t = \arg \max_{\theta_t} \sum_{i=1}^m \log c(F_t(y_{1it} | \hat{\delta}_{1t}), G_t(y_{2it} | \hat{\delta}_{2t}) | \theta_t).$$

We perform the IFM estimation method for the proposed bivariate copula-based linear mixed-effects models and time-varying bivariate copula-based linear models, by using the `optim` function (method “L-BFGS-B” by [2]) in R.

### 3.2 Interval Estimation

There are no analytical results readily available for obtaining the standard errors of the IFM estimates for the parameters of the bivariate copula-based linear mixed-effects models and time-varying bivariate copula-based linear models proposed in this paper. We then resorted to using a bootstrap resampling procedure in order to obtain confidence intervals for the models' parameters. However, the basic nonparametric bootstrap would not perform well here (mainly for the case of bivariate copula-based linear mixed-effects models) because of the increased dependence due to the presence of repeated observations in each of the resampled datasets. We chose, therefore, to run a parametric bootstrap procedure, which samples from the model using the (point) IFM estimates as the true values of the parameters. We performed a total of 1000 bootstrap samples for each proposed model. Then, we built 95% bootstrap confidence intervals by using the

percentile method by [9]. Besides its simplicity, which is the main attraction of this bootstrap method according to [3], no invalid parameter values can be included in the obtained intervals.

## 4 APPLICATION TO LONGITUDINAL CHILD GROWTH DATA

In this section, we show first the main results of an exploratory analysis of the child growth data (Section 4.1), which were drawn from a longitudinal study that was conducted with 150 children born in a maternity hospital in the municipality of Mutuípe, Bahia, from June 2005 to May 2006. Then, we present and discuss the main estimation results obtained via the univariate approach, i.e. with the separate linear mixed-effects models' fittings (Section 4.2), and also by the bivariate approach, i.e. through the proposed copula-based modelings (Section 4.3).

### 4.1 Descriptive Analysis

The children were measured monthly, from birth to six months of life. Information on weight (in g), height (in cm), hemoglobin level in the blood (in g/dl) and TDB (in months) were collected repeatedly. At first, an exploratory data analysis was performed to verify whether the children's height and weight are changing in a similar or different fashion. Thus, we observe, among others, that the average birth height was 48.3 cm, ranging from 42.1 cm to 52.5 cm, and the average birth weight was 3181.0 g, ranging from 2230.0 g to 4500.0 g. Moreover, in the last month of the study, about 25% of the children measured up to 64.1 cm and weighed up to 6909.0 g. The average follow-up time was 3.8 months, ranging from 2 to 6 months.

In studies with two longitudinal outcomes, it is useful to understand the strength of the relationship between them and the pattern of correlation across time.

Figure 1 shows the correlation between the children's height and weight over the 6 months period (corrgram). Among others, it suggests a joint analysis of both longitudinal outcomes (which are performed and discussed in Section 4.3), in order to understand growth and weight gain patterns over time.

As mentioned in Section 2.1, the choice of covariance matrices of the models was made on the basis of AIC values and by the construction of semi-variograms. Figure 2, left and right panels, show the associations between repeated observations, at time intervals, for height and weight, respectively. The horizontal line represents the estimation of the process variance. As can be seen from the right panel of Figure 2, the correlation decreases throughout the period considered, whereas from the left panel of this figure, we observe that the correlation first decreases, but it increases again in the last time points.

Furthermore, graphical methods can be used to explore the magnitude of between-person variability in outcomes over time. For instance, individual line plots for each study participant allow inspection of the individual response patterns and whether there is strong heterogeneity in the trajectories. Figure 3 shows the individual profile of the children's height (left panel) and weight (right panel). The black solid lines represent the mean response profile for height and weight,

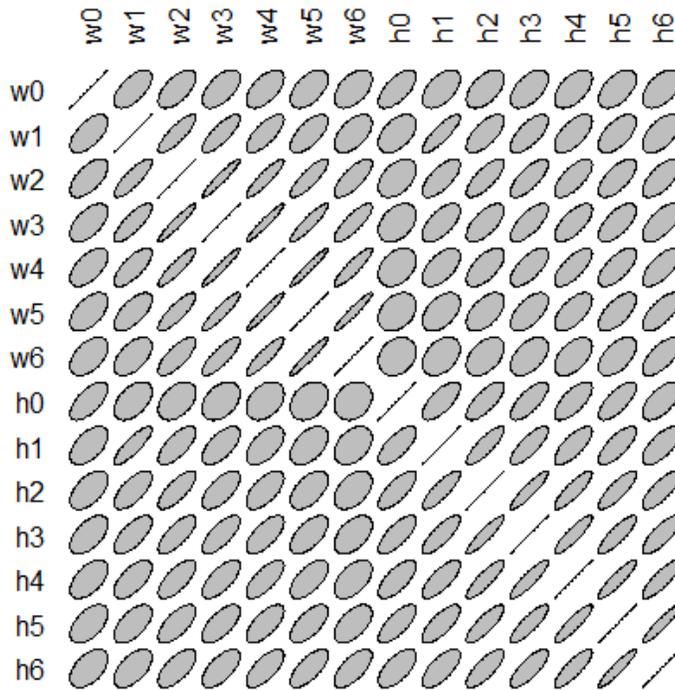


Figure 1: Corrogram for child growth data, where  $w_t$  and  $h_t$  denote, respectively, the weight and height at time  $t, t = 0, 1, \dots, 6$ .

against time/age (trend lines). Notice that, for both longitudinal outcomes, there exist a lower variability at the beginning and a greater variability at the end of the study (after some months of life), which is most notable for the weight variable. In addition to this, there is a greater within-person variability in the weight profile than in the height one, since the change in each individual line (slope) is more evident for the former than the latter. Thus, this descriptive analysis was essential for guiding our modeling strategy, indicating different trajectories with possibly different slopes for the weight outcome.

#### 4.2 Separate (Univariate) Linear Mixed-Effects Models

Table 3 shows the main estimation results for the linear mixed-effects models (2.2) and (2.3) when fitted separately (univariate analysis). It can be seen that the baby girls' growth is lower than that of baby boys (on average, 1.439 cm smaller). There is also a significant decrease of the hemoglobin level in the mean height corresponding to an increase in age (on average, 0.396 cm smaller). Finally, TDB does not contribute significantly to the growth pattern across time (p-value  $> 0.05$ ). The estimation results for model (2.3) are similar to those for model (2.2), in terms of the contribution and direction of the effect of each covariate. The only exception is the positive effect of TDB (at 5% level of significance), revealing a significant increase of the TDB in the mean

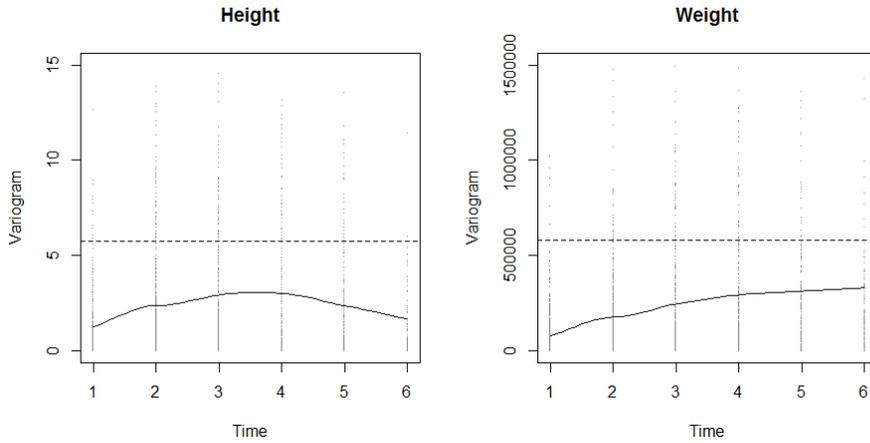


Figure 2: Semi-variogram for height and weight data.

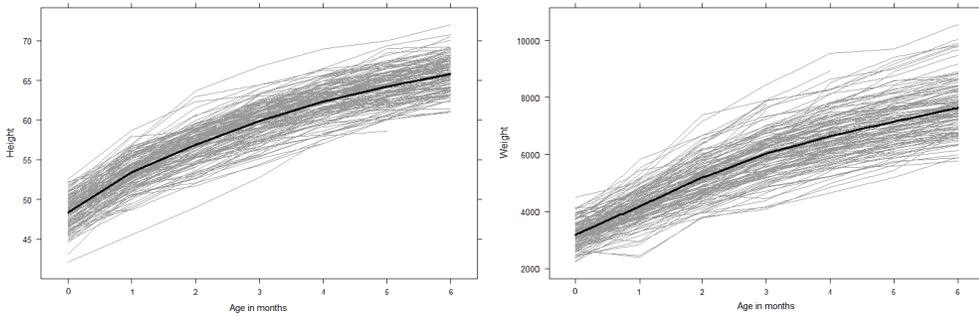


Figure 3: Longitudinal distribution of children’s height and weight. Mutuípe, Bahia, 2005-2006.

weight corresponding to an increase in age (on average, 59.228 g heavier). A random intercept for the growth pattern means that there is some average height in the population. Thus, the covariance structure considered in this analysis was constant and positive over time (compound symmetry or exchangeable correlation structure). For the weight model, the random intercept and age mean that the individual’s weight development varies not only in its initial stage, but its rate of change also varies because there are both intra- and inter-individual differences. The AIC and BIC (Bayesian Information Criterion; see [28]) values for each fitted model are also presented in Table 3. The overall AIC and BIC values, which are used for comparing the results from the univariate approach with the ones obtained from the bivariate approach (see Section 4.3), are calculated by simply summing their individual values, i.e.  $AIC_{Total} = 12140.400 + 2924.868 = 15065.268$  and  $BIC_{Total} = 12182.610 + 2962.385 = 15144.995$ .

We assessed the models’ adequacy by analyzing the conditional residuals of each fitted model. The conditional residual is given by the difference between the observed data and the predicted

Table 3: RML estimation results for the (univariate) linear mixed-effects models (2.2) and (2.3), fitted to the longitudinal child growth data. SE - standard error.

Fixed-Effect Parameter <sup>‡</sup>	Model (2.2): Height			Model (2.3): Weight		
	Value	SE	p-value	Value	SE	p-value
Intercept	55.097	0.614	< 0.0001	4495.871	156.308	< 0.0001
Gender (Female)	-1.439	0.307	< 0.0001	-286.155	77.529	0.0003
TDB	0.114	0.097	0.2415	59.229	24.142	0.0153
Age	2.612	0.035	< 0.0001	697.081	14.011	< 0.0001
Hemoglobin Level	-0.395	0.023	< 0.0001	-90.555	6.413	< 0.0001
Log-likelihood	-1454.434	—	—	-6061.202	—	—
AIC	2924.868	—	—	12140.400	—	—
BIC	2962.385	—	—	12182.610	—	—

<sup>‡</sup> Random effects of model (2.2):  $\hat{\sigma}_{\omega_{10}} = 1.401$ ;  $\hat{\rho}_1 = 0.665$ ;  $\hat{\sigma}_1 = 1.706$ .  
 Random effects of model (2.3):  $\hat{\sigma}_{\omega_{20}} = 414.382$ ;  $\hat{\sigma}_{\omega_{23}} = 141.663$ ;  $\hat{\rho}_2 = 0.227$ ;  $\hat{\sigma}_2 = 304.576$ .

value of the observation, i.e.  $e_{it} = Y_{it} - X'_{it}\hat{\beta} - Z'_{it}\hat{\omega}_i$ , for  $i = 1, \dots, m$  and  $t = 1, \dots, n_i$ , where  $\hat{\beta}$  and  $\hat{\omega}_i$  are, respectively, the RML estimates of  $\beta$  and  $\omega_i$ . Figures 4 and 5 show diagnostic plots (normal quantile plots of the standardized conditional residuals with simulated 95% confidence envelopes, and scatter-plots of the standardized conditional residuals versus fitted values) for models (2.2) and (2.3), respectively. The standardized conditional residual is defined as  $e^*_{it} = e_{it} / \sqrt{\hat{r}_{it}}$ , where  $\hat{r}_{it}$  is the  $(t, t)$ -th entry of  $\hat{R}_i$ , which is the RML estimate of  $R_i$ . These figures indicate the suitability of the proposed linear mixed-effects models for the height and weight outcomes, since we observe that the normality assumption of standardized conditional residuals is valid (see Figures 4 and 5's left panels) and, in general, there are no standardized conditional residuals with high values (see Figures 4 and 5's right panels).

### 4.3 Bivariate Copula-based Longitudinal Data Models

[15] proposed a graphic tool, called Kendall's plot (or K-plot), for detecting dependence in multivariate data. Figure 6 shows the Kendall's plot, per month, built from the longitudinal child growth data, which indicates a considerable positive (and of approximately the same magnitude) association between the height and weight outcomes.

Among the bivariate copula models that allow for positive association between the margins, we highlight the Gaussian, Student's  $t$  with arbitrary degrees of freedom (e.g.  $\nu = 2$ ), Clayton, Frank and Gumbel copulas. Table 4 exhibits the estimation results for these bivariate copula-based linear mixed-effects models fitted to the longitudinal Brazilian child growth data. The results include the 95% percentile bootstrap confidence intervals for the copula association parameter ( $\theta$ ), the corresponding measures of dependence (Kendall's tau, Spearman's rho, lower and upper tail dependence coefficients), and the AIC and BIC values. Here, we focus on the copula association parameter estimate ( $\hat{\theta}$ ), since the marginal parameter estimates are the same presented in Table 3 (we employed the IFM estimation method described in Section 3.1). It can be clearly seen from Tables 3 and 4 that the bivariate copula-based approach overcomes the univariate one (all

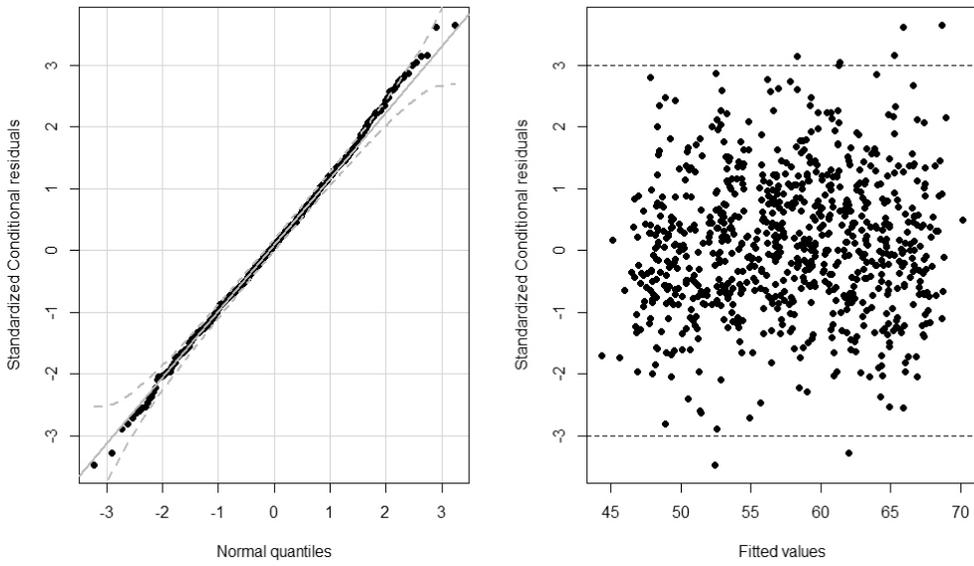


Figure 4: Diagnostic plots for the height model.

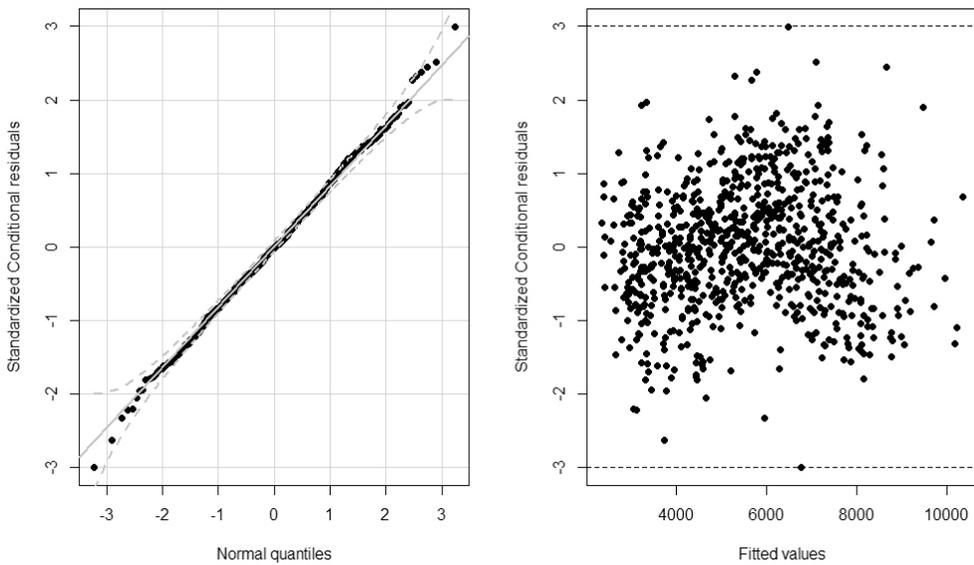


Figure 5: Diagnostic plots for the weight model.

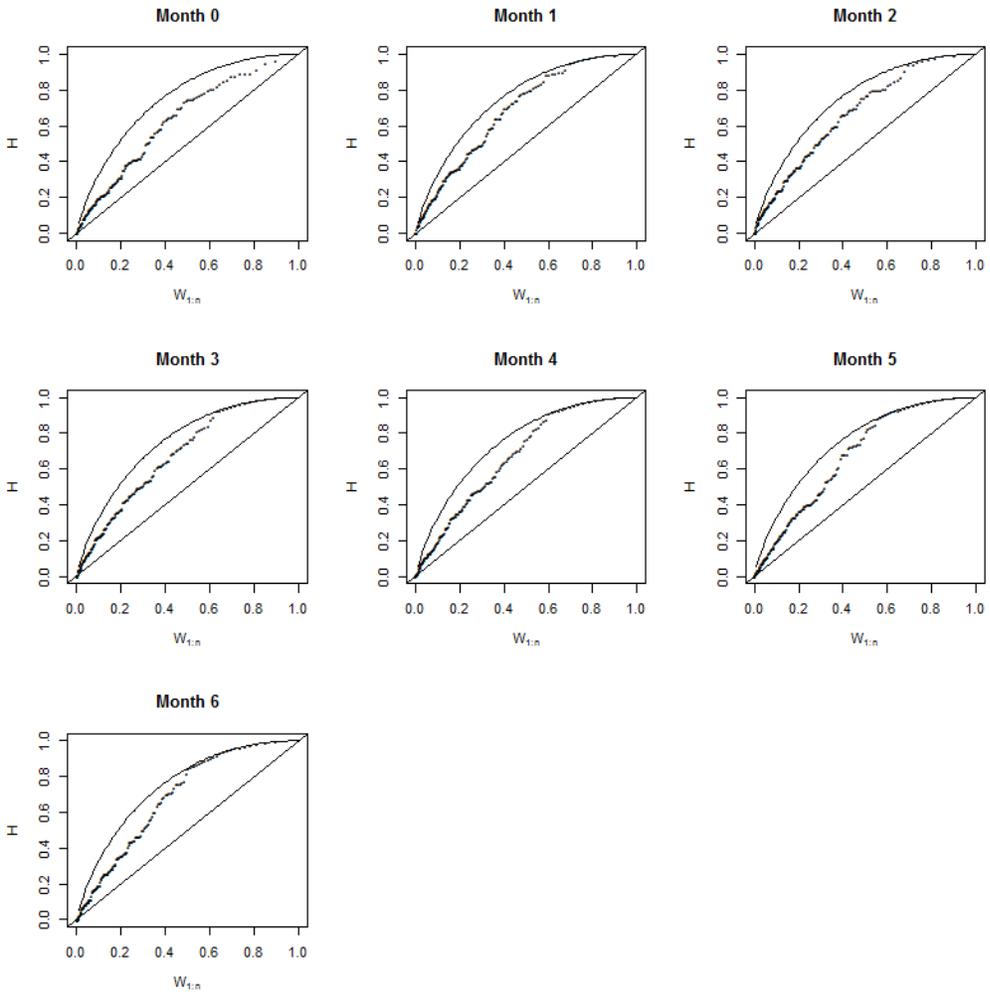


Figure 6: The Kendall's plot, per month, corresponding to the bivariate data at hand.

the copula-based models have smaller AIC and BIC values than the univariate models), justifying joint estimation of the height and weight's regression models through copulas to improve statistical efficiency. Table 4 also shows us that the bivariate linear mixed-effects model based on Gumbel copula was selected as the best one, according to both considered criteria (it has the smallest values on these criteria:  $AIC = 14429.37$  and  $BIC = 14513.89$ ). For this best model, the estimated dependence measures ( $\hat{\tau}_K = 0.789$  and  $\hat{\rho}_S = 0.937$ ) reveal a strong positive relationship between the margins (height and weight's regression models), and the estimated coefficient of tail dependence ( $\hat{\lambda}_U = 0.842$ ) shows the strongest positive relationship at the upper tail of the joint distribution, i.e. for high heights and weights.

Table 4: IFM estimation results for the bivariate copula-based linear mixed-effects models fitted to the data at hand, with a focus on the copula association parameter. LL - loglikelihood.

Copula	$\hat{\theta}$	SE	2.5%	97.5%	$\hat{\tau}_K$	$\hat{\rho}_S$	$\hat{\lambda}_L$	$\hat{\lambda}_U$	LL	AIC	BIC
Gaussian	0.920	0.044	0.784	0.950	0.743	0.912	0.000	0.000	-7198.01	14432.02	14516.54
Student's $t_2$	0.885	0.063	0.686	0.932	0.691	0.875	0.647	0.647	-7208.43	14452.86	14537.39
Clayton	2.634	0.668	1.222	3.851	0.568	0.755	0.768	0.000	-7214.14	14464.28	14548.81
Frank	15.046	2.768	8.219	18.584	0.763	0.929	0.000	0.000	-7197.90	14431.80	14516.32
Gumbel	4.747	0.891	2.620	6.118	0.789	0.937	0.000	0.842	-7196.68	14429.37	14513.89

Table 5 shows the IFM estimation results for the time-varying bivariate copula-based linear models' association parameters  $(\theta_0, \theta_1, \dots, \theta_6)$ , as well as their 95% percentile bootstrap confidence intervals (the results for the time-varying coefficients and standard deviation are not shown here). The estimated dependence measures for each of these models are presented in Table 6. It can be seen that the best bivariate model is the one based on Gaussian copula, which obtained the lowest AIC and BIC values (also presented in Table 5). Note, however, that it does not surpass the bivariate Gumbel copula-based linear mixed-effects model (with constant association parameter  $\theta$ ) selected before (see Table 4), according to both considered criteria. The reason for this may also be due to the fact that there is an intersection among the 95% percentile intervals for the time-varying bivariate Gaussian copula-based linear model's association parameters, indicating no gain in considering/assuming non-constant dependence between the children's height and weight over time. Furthermore, notice that, according to these criteria, none of the time-varying bivariate copula-based linear models overcome the linear mixed-effects models based on bivariate copulas with constant association parameter  $\theta$  (see Tables 4 and 5).

## 5 FINAL REMARKS AND FURTHER RESEARCH

Linear mixed-effects models are very popular in practice, since they are easy to handle and interpret. However, in this work such models have shown to be of limited use because of the impossibility of jointly modeling children's height and weight over time, also taking into account the correlation between the response variables.

There are also, on the other hand, some limitations to the use of copula models. The added complexity may render the estimation procedure more cumbersome, resulting in increased difficulty in finding convergent estimates. Moreover, although there are specific uses and guidelines for choosing between different types of copula models, an extra step of selecting between candidate models is made necessary when using the proposed approach.

Future research may focus on developing copula models even further. Novel strategies that incorporate cross-timed dependencies can be devised, in an attempt to model the dependence all past values of one variable has on the other variable, for instance. A regression model on the copula parameter may also provide further insight into the relationship within the data, providing information on how the association between variables is influenced by the covariates. In addition

Table 5: IFM estimates and 95% percentile bootstrap confidence intervals for the time-varying bivariate copula-based linear models' association parameters.

Copula	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	LL	AIC	BIC
Gaussian	0.696 [0.608, 0.774]	0.825 [0.767, 0.879]	0.759 [0.681, 0.835]	0.753 [0.665, 0.830]	0.683 [0.578, 0.781]	0.646 [0.524, 0.767]	0.684 [0.578, 0.773]	-7773.21	15700.42	15911.56
Student's $t_2$	0.620 [0.473, 0.736]	0.765 [0.657, 0.848]	0.656 [0.496, 0.774]	0.692 [0.534, 0.801]	0.548 [0.357, 0.703]	0.468 [0.250, 0.656]	0.550 [0.366, 0.691]	-7825.55	15805.09	16016.23
Clayton	1.354 [0.932, 1.862]	2.009 [1.390, 2.704]	1.490 [0.973, 2.138]	1.836 [1.238, 2.549]	1.233 [0.768, 1.783]	1.018 [0.612, 1.587]	1.204 [0.785, 1.768]	-7846.63	15847.26	16058.39
Frank	5.606 [4.375, 7.002]	8.785 [7.138, 10.724]	6.620 [5.189, 8.450]	6.751 [5.120, 8.717]	5.004 [3.612, 6.637]	4.558 [3.125, 6.216]	5.205 [3.848, 6.782]	-7799.50	15753.00	15964.13
Gumbel	1.850 [1.577, 2.209]	2.381 [2.005, 2.932]	2.030 [1.723, 2.495]	1.996 [1.651, 2.463]	1.746 [1.471, 2.102]	1.591 [1.354, 1.969]	1.725 [1.473, 2.116]	-7812.39	15778.77	15989.91

Table 6: Estimated dependence measures for the time-varying bivariate copula-based linear models.

Copula	Measure	Month ( $t$ )						
		0	1	2	3	4	5	6
Gaussian	$\hat{\tau}_K$	0.490	0.617	0.548	0.542	0.478	0.447	0.479
	$\hat{\rho}_S$	0.678	0.812	0.743	0.737	0.665	0.628	0.666
	$\hat{\lambda}_L$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$\hat{\lambda}_U$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Student's $t_2$	$\hat{\tau}_K$	0.425	0.554	0.455	0.486	0.369	0.310	0.370
	$\hat{\rho}_S$	0.601	0.749	0.638	0.674	0.530	0.451	0.532
	$\hat{\lambda}_L$	0.387	0.505	0.413	0.441	0.340	0.294	0.341
	$\hat{\lambda}_U$	0.387	0.505	0.413	0.441	0.340	0.294	0.341
Clayton	$\hat{\tau}_K$	0.403	0.501	0.426	0.478	0.381	0.337	0.375
	$\hat{\rho}_S$	0.568	0.684	0.596	0.658	0.540	0.483	0.533
	$\hat{\lambda}_L$	0.599	0.708	0.628	0.685	0.569	0.506	0.562
	$\hat{\lambda}_U$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Frank	$\hat{\tau}_K$	0.492	0.629	0.544	0.551	0.456	0.427	0.469
	$\hat{\rho}_S$	0.686	0.829	0.744	0.751	0.643	0.607	0.658
	$\hat{\lambda}_L$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$\hat{\lambda}_U$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gumbel	$\hat{\tau}_K$	0.459	0.580	0.507	0.498	0.427	0.371	0.420
	$\hat{\rho}_S$	0.635	0.768	0.691	0.681	0.595	0.525	0.586
	$\hat{\lambda}_L$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$\hat{\lambda}_U$	0.545	0.662	0.593	0.584	0.512	0.454	0.505

to this, other challenge to model dependence structure would be how to handle missing values without being forced to discard data.

## ACKNOWLEDGMENTS

The authors would like to thank the Editorial Board and the reviewers for their valuable comments and suggestions which helped to improve the manuscript.

**RESUMO.** Estudos longitudinais com múltiplas variáveis respostas são comuns na área de saúde pública e, conseqüentemente, métodos estatísticos adequados são requeridos quando há interesse em analisar a evolução temporal de uma ou mais variáveis resposta. Contudo, especificar a função de densidade conjunta de todas as variáveis respostas e a estrutura de correlação entre elas, bem como as dificuldades numéricas encontradas na inferência estatística quando a dimensão do problema aumenta, são os principais obstáculos dos procedimentos de modelagem multivariada. Como alternativas, neste artigo apresentamos duas propostas para lidar com dados longitudinais multivariados. Primeiramente, mostramos uma abordagem univariada, com modelos lineares mistos ajustados a cada uma

das variáveis respostas separadamente. Em seguida, apresentamos uma modelagem conjunta dessas variáveis, por meio do uso de funções cópula. Ambas as metodologias são aplicadas a um conjunto de dados reais bivariados referentes ao crescimento infantil de crianças brasileiras.

**Palavras-chave:** cópula bivariada, modelos lineares mistos, dados longitudinais de crescimento, dependência variante no tempo.

## REFERENCES

- [1] H. Akaike. On entropy maximization principle. In P.R. Krishnaiah (editor), “Applications of Statistics”. North-Holland, Amsterdam (1993), pp. 27–41.
- [2] R.H. Byrd, P. Lu, J. Nocedal & C.Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**(5) (1995), 1190–1208.
- [3] J. Carpenter & J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, **19** (2000), 1141–1164.
- [4] P. Catalano & L. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87** (1992), 651–658.
- [5] D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**(1) (1978), 141–151.
- [6] D.R. Cox & N. Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, **79**(3) (1992), 441–461.
- [7] P.J. Diggle, P.J. Heagerty, K.Y. Liang & S.L. Zeger. “Analysis of longitudinal data”, volume 25 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 2 ed. (2002), 379 pp.
- [8] F. Domma, S. Giordano & P.F. Perri. Statistical modeling of temporal dependence in financial data via a copula function. *Comm. Statist. Simulation Comput.*, **38**(3-5) (2009), 703–728.
- [9] B. Efron & R.J. Tibshirani. “An introduction to the bootstrap”, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York (1993), 436 pp.
- [10] P. Embrechts, A.J. McNeil & D. Straumann. Correlation and dependence in risk management: properties and pitfalls. In “Risk management: value at risk and beyond”. Cambridge Univ. Press, Cambridge (2002), pp. 176–223.
- [11] K.T. Fang, S. Kotz & K.W. Ng. “Symmetric multivariate and related distributions”, volume 36 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London (1990), 220 pp.
- [12] G. Fitzmaurice, M. Davidian, G. Verbeke & M. Geert. “Longitudinal Data Analysis. Handbooks of Modern Statistical Methods”. Chapman & Hall/CRC, New York (2009).
- [13] M.J. Frank. On the simultaneous associativity of  $F(x, y)$  and  $x + y - F(x, y)$ . *Aequationes Mathematicae*, **19**(2-3) (1979), 194–226.

- [14] E.W. Frees & P. Wang. Copula credibility for aggregate loss models. *Insurance: Mathematics & Economics*, **38**(2) (2006), 360–373.
- [15] C. Genest & J.C. Boies. Detecting dependence with Kendall plots. *The American Statistician*, **57**(4) (2003), 275–284.
- [16] C. Genest & A.C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12** (2007), 347–368.
- [17] R.V. Gueorguieva & A. Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, **96**(455) (2001), 1102–1112.
- [18] E.J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, **55** (1960), 698–707.
- [19] A. Ida, N. Ishimura & M. Nakamura. Note on the measures of dependence in terms of copulas. *Procedia Economics and Finance*, **14** (2014), 273–279.
- [20] H. Joe. “Dependence modeling with copulas”, volume 134 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, Boca Raton, FL (2014), 462 pp.
- [21] H. Joe & J. Xu. The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia (1996).
- [22] P. Lambert & F. Vandenhende. 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** (2002), 3197–3217.
- [23] R.C. Littell, J. Pendergast & R. Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19** (2000), 1793–1819.
- [24] X. Liu, M.J. Daniels & B. Marcus. Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association*, **104**(486) (2009), 429–438.
- [25] H. Manner & O. Reznikova. A survey on time-varying copulas: specification, simulations, and application. *Econometric Reviews*, **31**(6) (2012), 654–687.
- [26] R.B. Nelsen. “An introduction to copulas”. Springer Series in Statistics. Springer, New York, 2 ed. (2006), 269 pp.
- [27] R Core Team. R: A language and environment for statistical computing (version 3.3.2). R Foundation for Statistical Computing, Vienna, Austria (2016). URL <https://www.R-project.org/>.
- [28] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, **6**(2) (1978), 461–464.
- [29] B. Schweizer & E.F. Wolff. On nonparametric measures of dependence for random variables. *Ann. Statist.*, **9**(4) (1981), 879–885.
- [30] A. Sklar. Fonctions de répartition à  $n$  dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8** (1959), 229–231.

- [31] J. Sun, E.W. Frees & M.A. Rosenberg. Heavy-tailed longitudinal data modeling using copulas. *Insurance Mathematics & Economics*, **42** (2008), 817–830.
- [32] G. Verbeke, S. Fieuws, G. Molenberghs & M. Davidian. The analysis of multivariate longitudinal data: a review. *Statistical Methods in Medical Research*, **23**(1) (2014), 42–59.
- [33] B.T. West, K.B. Welch & A.T. Galecki. “Linear Mixed Models: A Practical Guide Using Statistical Software”. Chapman & Hall, Boca Raton, 2 ed. (2014).
- [34] J. Yan. Multivariate modeling with copulas and engineering applications. In H. Pham (editor), “Handbook in Engineering Statistics”. Springer-Verlag, New York (2006), pp. 973–990.