The Canadian Journal of Statistics | La revue canadienne de statistique
Vol. 000, No. 00, 2024, e11830

A new copula regression model for hierarchical data

Talagbe Gabin AKPO
Université Laval & Institut national de la recherche scientifique (INRS), Centre Armand-Frappier Santé Biotechnologie, Laval, Québec H7V 1B7, Canada

Louis-Paul RIVEST
Department of Mathematics and Statistics, Université Laval, Québec, G1V 0A6, Canada

Abstract This article proposes multivariate copula models for hierarchical data. They account for two types of correlation: one is between variables measured on the same unit, and the other is a correlation between units in the same cluster. This model is used to carry out copula regression for hierarchical data that gives cluster-specific prediction curves. In the simple case where a cluster contains two units and where two variables are measured on each one, the new model is constructed with a D-vine. The proposed copula density is expressed in terms of three copula families. When the copula families and the marginal distributions are normal, the model is equivalent to a normal linear mixed model with random cluster-specific intercepts. Methods to select the three copula families and to estimate their parameters are proposed. We perform Monte Carlo studies of the sampling properties of these estimators and of out-of-sample predictions. The new model is applied to a dataset on the marks of students in several schools.

Résumé Cet article propose des modèles de copules pour données hiérarchiques. Ils tiennent compte de deux types de corrélation: l'une est entre les variables mesurées sur une même unité et l'autre est entre les unités d'une même grappe. Ce modèle est utilisé pour faire de la régression avec des copules qui donne des courbes de prédiction spécifiques à chaque grappe. Dans le cas simple où une grappe contient deux unités, le nouveau modèle est construit à l'aide d'une D-vigne. La densité de la nouvelle copule fait intervenir trois familles de copules. Quand elles sont toutes normales et que les marges sont normales, le modèle est équivalent à un modèle linéaire mixte normal avec ordonnées à l'origine aléatoires. Des méthodes pour choisir les trois familles de copules et pour estimer leurs paramètres sont mises de l'avant. On étudie, par simulation Monte-Carlo, les propriétés échantillonnales des estimateurs et la prédiction de nouvelles observations. Un jeu de données sur les notes d'écoliers dans plusieurs écoles est analysé avec le nouveau modèle.

Keywords Exchangeability; heterogeneity; normal linear mixed models; predictions; vine copula.
MSC2020 Primary 62J02; Secondary 62F99.
@ Corresponding author Louis-Paul.Rivest@mat.ulaval.ca

1 Introduction

A simple bivariate copula regression predicts a dependent variable Y using an independent variable X. It proceeds by selecting a bivariate copula summarizing the relationship between X and Y. The copula regression predictor is then constructed using characteristics of the conditional distribution of Y, given X, derived from the selected copula as discussed in Nelsen (2006, p. 217). Kumar and Shoukri (2007) and Crane and Van der Hoek (2008) exemplify this method, while Noh et al. (2013) derive its asymptotic properties and Acar et al. (2019) investigate prediction errors. Recent multivariate extensions can be found in Côté et al. (2019) and Hofert et al. (2023).

DOI: 10.1002/cjs.11830

The goal of this article is to generalize the basic copula regression model to hierarchical data: X and Y are observed on units that are in clusters, and one would like to include a cluster effect in the copula regression predictions. The classical regression model for hierarchical data is a normal linear mixed model, with cluster-specific random slopes and intercepts, as discussed in Battese et al. (1988), Verbeke and Molenberghs (2000) and McCulloch and Searle (2004). Rivest et al. (2016) and Grover et al. (2020) propose using exchangeable copula families to model the residual dependency within clusters in this context. Su et al. (2019) provide a survival data application of this approach.

A joint distribution of all the X and Y variables in a cluster is first constructed. The conditions that the model must fulfill in order to yield suitable predictions are given in Section 2. An exchangeability assumption is required: permuting the units in a cluster does not change the joint distribution of the variables. It also relies on a partial conditional independence assumption, which ensures that the prediction of Y for a unit does not depend on the X values for the other units in the cluster. The proposed model is then constructed with a D-vine in the simple case of a cluster containing two units. The general model, for clusters of arbitrary size, is introduced in Section 3. We use the proposed copula to make cluster-specific predictions. The new model is then applied to a dataset of Goldstein (2011); the results of this analysis are compared with those obtained with standard normal mixed linear models.

2 Model construction: Exchangeability and conditional independence

Throughout this article, $i \in \{1, \dots, m\}$ indexes clusters in the population, where m is the number of clusters. Subscript j, $j \in \{1, \dots, n_i\}$, represents a unit within cluster i whose size is n_i. Figure 1 presents a small population with $m = 3$ clusters of respective sizes 3, 5 and 4. The clusters are assumed to be independent. This section describes the joint distribution of the variables $(X_{ij}, Y_{ij})^{\mathrm{T}}$ observed on the n_i units of cluster i. To simplify the notation, subscript i is dropped and a joint distribution for the n pairs $\mathbf{Z}_j = (X_j, Y_j)^{\mathrm{T}}$ is proposed, where Y is the dependent variable in the regression while X is the explanatory variable.

Let $F_{2,1:n}(\mathbf{z}_1, \dots, \mathbf{z}_n)$ be the joint cumulative distribution function (cdf) of the $2n$ continuous variables measured in the cluster. The model is constructed using copulas; it is therefore of interest to define $F(x)$ and $G(y)$ as the marginal distributions of, respectively, X_j and Y_j, which are the same for the n units. We let $U_j = F(X_j)$ and $V_j = G(Y_j)$ be random variables with uniform margins, and $c_{2,1:n}(u_1, v_1, \dots, u_n, v_n)$ be the copula density for the joint distribution of $\{(U_j, V_j) : j = 1, \dots, n\}$. We now give some conditions for the family of joint distributions $F_{2,1:n}(\mathbf{z}_1, \dots, \mathbf{z}_n)$, $n \in \{2, 3, \dots\}$ to give useful regression models.

2.1 Exchangeability

We consider the family of cdfs defined by

$$\mathcal{F}_2 = \left\{F_{2,1:n}(\mathbf{z}_1, \dots, \mathbf{z}_n) : \mathbf{z}_j \in \mathbb{R}^2,\ j = 1, \dots, n,\ n = 1, 2 \dots\right\}.$$

Definition 1. The family $\mathcal{F}_2$ is said to be 2-exchangeable if, for all $n \geq 2$, $F_{2,1:n} \in \mathcal{F}_2$ satisfies the following conditions:

(i) Permutation invariance: For all permutations $\{\pi(1), \dots, \pi(n)\}$ of $\{1, 2, \dots, n\}$,

$$F_{2,1:n}(\mathbf{z}_1, \dots, \mathbf{z}_n) = F_{2,1:n}\left(\mathbf{z}_{\pi(1)}, \dots, \mathbf{z}_{\pi(n)}\right). \tag{1}$$

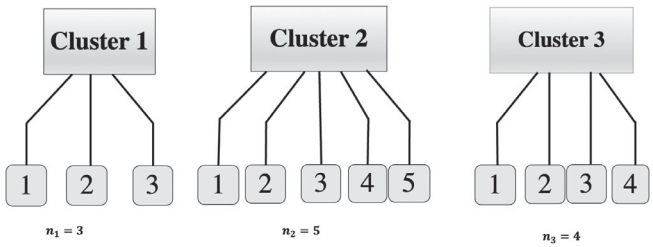

Figure 1: A small population with three clusters of sizes 3, 5 and 4.

(ii) Closure on marginalization: For any $r \leq n$,

$$F_{2,1:r}(\mathbf{z}_1, \ldots, \mathbf{z}_r) = F_{2,1:n}(\mathbf{z}_1, \ldots, \mathbf{z}_r, \infty, \ldots, \infty). \tag{2}$$

Equation (1) gives the classical definition of univariate exchangeability, or 1-exchangeability, which is discussed in Mai and Scherer (2012, chap. 1). A simple example of 2-exchangeability is a bivariate one-way ANOVA model with random effects, $\mathbf{Z}_j = \mathbf{A} + \mathbf{E}_j$, where $\mathbf{A}$ and $\mathbf{E}_j$, $j \in \{1, \ldots, n\}$, are $n+1$ independent random vectors of dimension 2 and the $\mathbf{E}_j$'s have identical distributions. The next proposition, whose proof is given in the Supplementary Material, gives the form of the correlation matrix for a 2-exchangeable random vector.

Proposition 1. *Let $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ be a set of 2×1 random vectors of size n whose joint distribution is 2-exchangeable. Then, the $2n \times 2n$ matrices of Pearson, Spearman, and Kendall's correlations between these $2n$ variables have the form*

$$I_n \otimes \Sigma_w + J_n \otimes \Sigma_b = \begin{pmatrix} \Sigma_w + \Sigma_b & \Sigma_b & \ldots & \Sigma_b \\ \Sigma_b & \Sigma_w + \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \ldots & \Sigma_b & \Sigma_w + \Sigma_b \end{pmatrix}, \tag{3}$$

where J_n is an $n \times n$ matrix of 1s, $\otimes$ denotes the Kronecker product and Σ_b and Σ_w are symmetric 2×2 matrices. Moreover, for Pearson and Spearman correlations, the matrices Σ_w and Σ_b are positive semidefinite.

The definition of exchangeability given in this section is easily extended to an arbitrary dimension d. As this article focuses on a univariate explanatory variable X, it uses this notion only when $d = 2$.

2.2 Partial conditional independence

This section proposes an assumption concerning the dependency within each cluster. The random variables Y_j and $\{X_k : k \neq j\}$ are assumed to be independent, given X_j. This is a partial conditional independence assumption that can be written as

$$Y_j \perp \{X_k : k \neq j\} \mid X_j, \ j \in \{1, \ldots, n\}. \tag{4}$$

This condition is weaker than the conditional independence assumption underlying the standard regression model that can be formulated as

$$Y_j \perp \{(X_k, Y_k)^{\mathrm{T}} : k \neq j\} \mid X_j, \ j \in \{1, \ldots, n\}.$$

We now implement the 2-exchangeability and the partial independence assumption within a D-vine for the joint distribution of two units in a cluster.

2.3 A D-vine construction of the model when $n = 2$

This section constructs a copula density function for the random variables in a cluster containing two units and verifies the properties of partial conditional independence and of 2-exchangeability. The four random variables are $U_1 = F(X_1)$, $V_1 = G(Y_1)$, $U_2 = F(X_2)$ and $V_2 = G(Y_2)$. The three trees of the proposed D-vine are given in Figure 2.

The decomposition involves six bivariate copulas $C_{U_1V_1}$, $C_{U_1U_2}$, $C_{U_2V_2}$, $C_{V_1U_2;U_1}$, $C_{U_1V_2;U_2}$ and $C_{V_1V_2;U_1U_2}$, where the last three distributions are conditional on U_1, U_2 and (U_1, U_2), respectively. The density of the multivariate copula corresponding to the trees given in Figure 2, $c_{2,1:2}\{(u_1, v_1), (u_2, v_2)\}$, can be derived from Joe (2014, p. 108). It is given by

$$\begin{aligned} &c_{U_1V_1}(u_1, v_1) c_{U_2V_2}(u_2, v_2) c_{U_1U_2}(u_1, u_2) \times c_{V_1U_2;U_1}\{C_{V_1|U_1}(v_1|u_1), C_{U_2|U_1}(u_2|u_1)\} \times \\ &c_{U_1V_2;U_2}\{C_{U_1|U_2}(u_1|u_2), C_{V_2|U_2}(v_2|u_2)\} \times c_{V_1V_2;U_1U_2}\{C_{V_1|U_1U_2}(v_1|u_1, u_2), C_{V_2|U_1U_2}(v_2|u_1, u_2)\}, \end{aligned} \tag{5}$$

where the conditional distribution functions $C_{V|U}$ and $C_{W|UV}$ are defined by

$$C_{V|U}(v|u) = \frac{C_{UV}(u, v)}{\partial u}, \quad C_{W|UV}(w|u, v) = \frac{\partial C_{WV;U}\{C_{V|U}(v|u), C_{W|U}(w|u)\}}{\partial C_{V|U}(v|u)}, \tag{6}$$

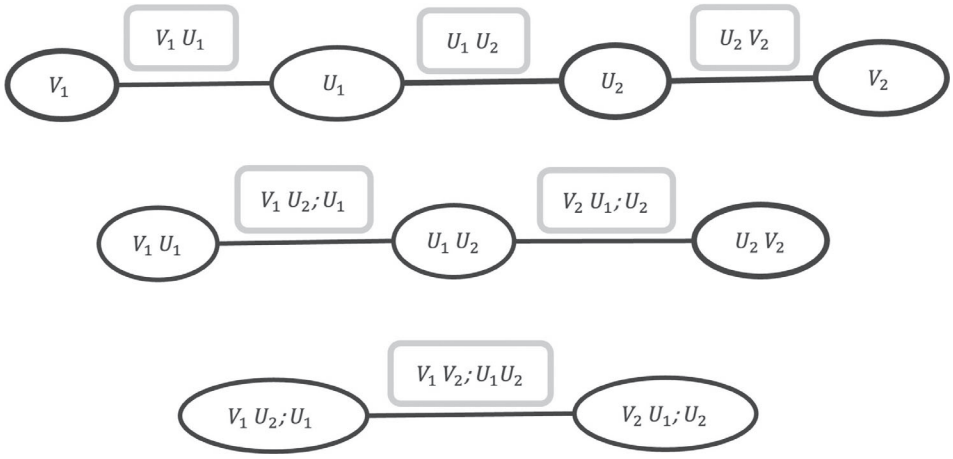

Figure 2: Graphical representation of the proposed D-vine.

and ∂ stands for the differentiation operator. This vine decomposition is discussed in Dissmann et al. (2013) and Czado and Nagler (2022). We would like the density in (5) to fulfill the conditions for 2-exchangeability and for partial conditional independence.

The density of the copula is 2-exchangeable if $c_{2,1:2}\{(u_1,v_1),(u_2,v_2)\} = c_{2,1:2}\{(u_2,v_2),(u_1,v_1)\}$. This requires a unique copula density for the dependency between U and V, that is, $c_{U_1V_1}(u,v) = c_{U_2V_2}(u,v)$, for $u,v \in (0,1)$. Also, the copula density for the dependency between U_1 and U_2 needs to be symmetric, that is, $c_{U_1U_2}(u,v) = c_{U_1U_2}(v,u)$ and $c_{V_1V_2;U_1U_2}(u,v) = c_{V_1V_2;U_1U_2}(v,u)$. This definition also entails restrictions on the conditional copula densities $c_{V_1U_2|U_1}$ and $c_{U_1V_2|U_2}$. However, for the D-vine in (5) to fulfill the partial conditional independence condition in (4), these two copulas need to be the independence copula. This leads to the density

$$c_{2,1:2}\{(u_1,v_1),(u_2,v_2)\} = c_{U_1U_2}(u_1,u_2)\ c_{U_1V_1}(u_1,v_1)\ c_{U_1V_1}(u_2,v_2)\ \ c_{V_1V_2;U_1U_2}\{C_{V_1|U_1}(v_1|u_1), C_{V_1|U_1}(v_2|u_2)\} \quad (7)$$

which involves an arbitrary copula density c_{UV} for the relationship between the explanatory variable X and the response Y, and two symmetric copula densities $c_{U_1U_2}$ and $c_{V_1V_2;U_1U_2}$ for the dependency between units.

The next proposition considers the special case where the three copulas in (7) are normal. Their density depends on a correlation $\rho \in (-1,1)$; it is given by

$$c(u_1,u_2) = \frac{\exp[-\{\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2 - 2\rho\Phi^{-1}(u_1)\Phi^{-1}(u_2)\}/\{2(1-\rho^2)\}]}{\sqrt{1-\rho^2}\exp[-\{\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2\}/2]}, \quad u_1,u_2 \in (0,1).$$

Let ρ_{U_1,U_2} represent the correlation between U_1 and U_2, ρ_{U_1,V_1} that between U_j and V_j for $j = 1,2$, and $\rho_3 = \rho_{V_1,V_2|U_1,U_2}$ be the residual correlation of V_1 and V_2 given U_1 and U_2.

Proposition 2. *When the three copulas of the decomposition given in (7) are normal, the joint copula for (U_1,U_2,V_1,V_2) is normal with a correlation matrix given by*

$$\begin{pmatrix} 1 & \rho_{U_1,U_2} & \rho_{U_1,V_1} & \rho_{U_1,U_2}\rho_{U_1,V_1} \\ \rho_{U_1,U_2} & 1 & \rho_{U_1,U_2}\rho_{U_1,V_1} & \rho_{U_1,V_1} \\ \rho_{U_1,V_1} & \rho_{U_1,U_2}\rho_{U_1,V_1} & 1 & \rho_{U_1,U_2}\rho^2_{U_1,V_1} + \rho_3(1-\rho^2_{U_1,V_1}) \\ \rho_{U_1,U_2}\rho_{U_1,V_1} & \rho_{U_1,V_1} & \rho_{U_1,U_2}\rho^2_{U_1,V_1} + \rho_3(1-\rho^2_{U_1,V_1}) & 1 \end{pmatrix}. \quad (8)$$

Proof. The joint copula density (7) for (U_1,U_2,V_1,V_2) is normal (Joe, 2014, p. 119). To derive the correlation matrix in (8), the result that $\mathrm{E}\{\Phi^{-1}(U_1)\Phi^{-1}(V_2)\} = \rho_{U_1,U_2}\rho_{U_1,V_1}$ is obtained by conditioning on U_2, while the variance-covariance matrix of $\{\Phi^{-1}(V_1),\Phi^{-1}(V_2)\}$ knowing (U_1,U_2) is used to find that $\rho_3 = [\mathrm{E}\{\Phi^{-1}(V_1)\Phi^{-1}(V_2)\} - \rho_{U_1,U_2}\rho^2_{U_1,V_1}]/(1-\rho^2_{U_1,V_1})$. □

Suppose that the marginal distributions of X_1 and X_2 are $\mathcal{N}(\mu_1,\sigma_1^2)$, while those for Y_1 and Y_2 are $\mathcal{N}(\mu_2,\sigma_2^2)$ and that the copula for (X_1,X_2,Y_1,Y_2) is normal with a correlation matrix given in (8). The relationship

between X and Y can then be expressed using the following normal mixed linear model:

$$Y_j = \beta_0 + \beta_1 X_j + \sigma_2\sqrt{\rho_3}(1-\rho^2_{U_1,V_1})^{1/2}A + \sigma_2(1-\rho^2_{U_1,V_1})^{1/2}(1-\rho_3)^{1/2}E_j, \quad j = 1,2, \tag{9}$$

where $\beta_1 = \sigma_2\rho_{U_1,V_1}/\sigma_1$, $\beta_0 = \mu_2 - \beta_1\mu_1$ and A, E_1, E_2 are independent random variables with a $\mathcal{N}(0,1)$ distribution. This is the random intercept linear mixed model of Battese et al. (1988), labelled $ML1$. One can easily show that the correlation matrix of (X_1, X_2, Y_1, Y_2) defined in (9) is given by (8). This model involves three variances: that of Y given X, one for the random intercept, $\sigma_2\sqrt{\rho_3}(1-\rho^2_{U_1,V_1})^{1/2}A$, and one for the experimental errors, $\sigma_2(1-\rho^2_{U_1,V_1})^{1/2}(1-\rho_3)^{1/2}E_j$, for $j = 1,2$. These three variances are, respectively, given by

$$\sigma_r^2 = \sigma_2^2(1-\rho^2_{U_1,V_1}), \quad \sigma_v^2 = \rho_3\sigma_r^2, \quad \sigma_e^2 = (1-\rho_3)\sigma_r^2. \tag{10}$$

Note also that this model is easily generalized to clusters of size $n > 2$. This is also true of the density in (7) as shown in the next section.

3 A multivariate 2-exchangeable copula model

This section proposes a method to construct densities for the joint distribution of the variables in a cluster that meet the constraints of 2-exchangeability and of partial conditional independence presented in Section 2.

3.1 Model construction

As discussed at the beginning of Section 2, the joint distribution $F_{2;1:n}$ of $\mathbf{Z}_i, i \in \{1,\dots,n\}$, the 2×1 vectors observed on the n units of a cluster, is expressed in terms of the marginal distributions $F(x)$ and $G(y)$ and of a 2-exchangeable copula. The density of this copula is constructed using three components: $\{c^{(1)}_{1,1:n}(u_1,\dots,u_n) : n \geq 1\}$, a 1-exchangeable family of copula densities for the dependence of the X-variables in a cluster; $c^{(2)}(u,v)$, a bivariate copula density for the marginal relationship between X and Y; and $\{c^{(3)}_{1,1:n}(w_1,\dots,w_n) : n \geq 1\}$, a 1-exchangeable family of copula densities for the residual dependency. The proposed 2-exchangeable copula density is

$$c_{2,1:n}(u_1,v_1,\dots,u_n,v_n) = c^{(1)}_{1,1:n}(u_1,\dots,u_n)\prod_{j=1}^{n}\{c^{(2)}(u_j,v_j)\} \times c^{(3)}_{1,1:n}\left\{C^{(2)}_{2|1}(v_1|u_1),\dots,C^{(2)}_{2|1}(v_n|u_n)\right\}, \tag{11}$$

where $u_j, v_j \in [0,1]$, $j = 1,\dots,n$ and the conditional distribution $C^{(2)}_{2|1}$ is deduced from (6) with C_{UV} replaced by $C^{(2)}$. When $n = 2$, (11) reduces to the D-vine copula density in (7).

Proposition 3. *Define, for $x_j, y_j \in \mathbb{R}$, $j \in \{1,\dots,n\}$,*

$$f_{2,1:n}(x_1,y_1,\dots,x_n,y_n) = \left\{\prod_{j=1}^{n} f(x_j)g(y_j)\right\} \times c_{2,1:n}\{F(x_1),G(y_1),\dots,F(x_n),G(y_n)\},$$

where $f(x)$ and $g(y)$ are the densities for $F(x)$ and $G(y)$, and $c_{2,1:n}$ is given by formula (11). Then, $f_{2,1:n}$ is the density of $2n$ random variables (X_j, Y_j), $j \in \{1,\dots,n\}$, and it satisfies the conditions (1) and (2) for 2-exchangeability and the partial conditional independence assumption in (4).

Proof. Without loss of generality, we assume that the marginal distributions $F(x)$ and $G(y)$ are uniform on $(0,1)$; thus we need to show that the function given in (11) is a copula density. First we integrate $c_{2,1:n}$ for $v_n \in (0,1)$. To carry this out, it is convenient to make the change of variable, $w_n = C^{(2)}_{2|1}(v_n|u_n)$. The Jacobian is $dw_n = c^{(2)}(u_n,v_n)dv_n$. Using the closure on marginalization property of the family $C^{(3)}_{1,1:n}$, the integral is equal to

$$c^{(1)}_{1,1:n}(u_1,\dots,u_n)\left\{\prod_{j=1}^{n-1} c^{(2)}(u_j,v_j)\right\}c^{(3)}_{1,1:(n-1)}\left\{C^{(2)}_{2|1}(v_1|u_1),\dots,C^{(2)}_{2|1}(v_{n-1}|u_{n-1})\right\}.$$

The variable u_n appears only in $c^{(1)}_{1,1:n}(u_1,\dots,u_n)$. Using the closure on marginalization property of the family $C^{(1)}_{1,1:n}$, the integral on u_n gives the density (11) for the $(n-1)$ pairs $(U_1,V_1),\dots,(U_{n-1},V_{n-1})$. Thus expression (11) defines a proper copula density that meets the requirements in (1) and (2) for 2-exchangeability. To prove the partial conditional independence assumption in (4), one integrates the above density for $v_{n-1},v_{n-2},\dots,v_2\in(0,1)$. This is easily carried out by changing variables, $w_j=C^{(2)}_{2|1}(v_j|u_j),\ j\in\{2,\dots,n-1\}$. The joint density of $(V_1,U_1,\dots,U_n)$ is found to be $c^{(1)}_{1,1:n}(u_1,\dots,u_n)c^{(2)}(u_1,v_1)$; thus, given U_1,V_1 and $(U_2,\dots,U_n)$ are independent and the partial conditional independence assumption holds. □

The derivations in the previous paragraph have highlighted a key property of the proposed model. If the density of $(U_1,V_1,\dots,U_n,V_n)$ is given by expression (11), then the two vectors $(U_1,\dots,U_n)$ and $\{W_1=C_{2|1}(V_1|U_1),\dots,W_n=C_{2|1}(V_n|U_n)\}$ are independent with densities, respectively, given by $c^{(1)}_{1,1:n}(u_1,\dots,u_n)$ and $c^{(3)}_{1,1:n}(w_1,\dots,w_n)$. This is summarized in the following proposition.

Proposition 4. *Let $(U_1,V_1,\dots,U_n,V_n)$, be a set of $2n$ random variables, whose joint density is given by expression (11). If $W_j=C^{(2)}_{2|1}(V_j|U_j),\quad j\in\{1,\dots,n\}$, where the distribution function $C^{(2)}_{2|1}$ comes from (6), then the random vectors $(U_1,\dots,U_n)$ and $(W_1,\dots,W_n)$ are independent with respective densities $c^{(1)}_{1,1:n}$ and $c^{(3)}_{1,1:n}$.*

Proposition 4 suggests a simple algorithm to simulate a random vector with a 2-exchangeable copula density given by (11). First, simulate $(U_1,\dots,U_n)$ and $(W_1,\dots,W_n)$ according to independent 1-exchangeable copulas $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$, respectively, and then solve $W_j=C^{(2)}_{2|1}(V_j|U_j), j=1,\dots,n$, to get $(V_1,\dots,V_n)$. As seen in Section 2, the 2-exchangeable copula model can be constructed using an R-vine when $n=2$. The Supplementary Material shows that this remains true when $n>2$ as long as the copulas $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$ can be decomposed using D-vines.

3.2 The special case where $C^{(2)}$ is a normal copula

We now assume that the copula $C^{(2)}$ is bivariate normal with correlation ρ_2. The conditional distribution of V, given U, and its inverse are given by

$$C^{(2)}_{2|1}(v|u)=\Phi\left\{\frac{\Phi^{-1}(v)-\rho_{U_1,V_1}\Phi^{-1}(u)}{\sqrt{1-\rho^2_{U_1,V_1}}}\right\}\text{ and }\{C^{(2)}_{2|1}\}^{-1}(t|u)=\Phi\left\{\Phi^{-1}(t)\sqrt{1-\rho^2_{U_1,V_1}}+\rho_{U_1,V_1}\Phi^{-1}(u)\right\},$$

for $u,v,t\in[0,1]$ (Bernard and Czado, 2015). Using Proposition 4, the dependent variable Y_j with explanatory variable $X_j=x_j$ can be expressed as $Y_j=G^{-1}[\{C^{(2)}_{2|1}\}^{-1}\{W_j|F(x_j)\}]$, where W_j is the entry of a vector with distribution $C^{(3)}_{1,1:n}$. This gives the following linear model:

$$\Phi^{-1}\{G(Y_j)\}=\rho_{U_1,V_1}\Phi^{-1}\{F(x_j)\}+\sqrt{1-\rho^2_{U_1,V_1}}\Phi^{-1}(W_j),\ j=1,\dots,n$$

for the units in a cluster. If, in addition, the marginal distributions of X and Y are normal with respective means μ_1,μ_2 and variances (σ_1^2,σ_2^2), as discussed in (9), then the model becomes $Y_j=\beta_0+\beta_1x_j+\sigma_r\Phi^{-1}(W_j),\ j=1,\dots,n$, where β_0, β_1 and σ_r^2 are defined in (9) and (10). Then the marginal conditional distributions of Y_j, given X_j, are normal. However, their joint conditional distribution depends on the copula $C^{(3)}_{1,1:n}$. Indeed, the conditional density for $(Y_1,\dots,Y_n)$, given $(X_1=x_1,\dots,X_n=x_n)$, is equal to

$$f(y_1,\dots,y_n|x_1,\dots,x_n)=\frac{1}{(2\pi)^{n/2}\sigma_r^n}\exp\left\{-\sum_{j=1}^{n}\frac{(y_i-\beta_0-\beta_1x_j)^2}{2\sigma_r^2}\right\}\times$$
$$c^{(3)}_{1,1:n}\left[\Phi\left\{\frac{y_1-\beta_0-\beta_1x_1}{\sigma_r}\right\},\dots,\Phi\left\{\frac{y_n-\beta_0-\beta_1x_n}{\sigma_r}\right\}\right]\quad y_j\in\mathbb{R},j\in\{1,\dots,n\}.$$

3.3 Predictions with the 2-exchangeable model

Suppose that $(n-1)$ units $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$ have been observed in a cluster and that only the covariate x_n is known for unit n. This section investigates the prediction of Y_n given x_n and $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$. An expression for the conditional expectation of Y_n given these variables is derived. Illustrations of the prediction curves for various specifications of the copula $C^{(2)}$ are presented.

The next proposition gives an expression for $g_P(y_n)$, the conditional density of Y_n, given x_n and $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$.

Proposition 5. *Let* $w_j = C^{(2)}_{2|1}\{G(y_j)|F(x_j\}, j \in \{1, \dots, n-1\}$ *and* $W_n = C^{(2)}_{2|1}\{G(Y_n)|F(x_n)\}$. *Then the conditional density of* Y_n, *given* x_n *and* $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$, *is given by*

$$g_P(y_n) = g(y_n)c^{(2)}\{F(x_n), G(y_n)\}\frac{c^{(3)}_{1,1:n}\left[w_1, \dots, w_{n-1}, C^{(2)}_{2|1}\{G(y_n)|F(x_n)\}\right]}{c^{(3)}_{1,1:(n-1)}(w_1, \dots, w_{n-1})} \quad y_n \in \mathbb{R}.$$

In addition, the conditional expectation of Y_n, given x_n and $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$, is

$$\mathrm{E}_P(Y_n) = \int_0^1 G^{-1}\left[\left\{C^{(2)}_{2|1}\right\}^{-1}\{w|F(x_n)\}\right]\frac{c^{(3)}_{1,1:n}(w_1, \dots, w_{n-1}, w)}{c^{(3)}_{1,1:(n-1)}(w_1, \dots, w_{n-1})}dw. \tag{12}$$

Proof. Dividing the density in (11) by the joint density for $(X_1, Y_1, \dots, X_{n-1}, Y_{n-1}, X_n)$ derived in Proposition 3 gives the formula $g_P(y_n)$. Changing the variable $W_n = C_{2|1}\{G(Y_n)|F(x_n)\}$ in this density shows that the conditional density of W_n, given $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$, is simply $c^{(3)}_{1,1:n}(w_1, \dots, w_n)/c^{(3)}_{1,1:(n-1)}(w_1, \dots, w_{n-1})$. This is the conditional distribution of $(W_n|W_1, \dots, W_{n-1})$, when $(W_1, \dots, W_n)$ is distributed according to copula $C^{(3)}_{1,1:n}$. The best predictor of the unknown Y_n is the conditional expectation of $G^{-1}[\{C^{(2)}_{2|1}\}^{-1}\{W_n|F(x_n)\}]$, which is given by Equation (12). □

If the copula $C^{(3)}_{1,1:n}$ is the independence copula, then Equation (12) reduces to a standard unconditional copula regression for $C^{(2)}$. Taking the expectation of $\mathrm{E}_P(Y_n)$ with respect to the distribution of $(X_1, Y_1, \dots, X_{n-1}, Y_{n-1})$ also gives the unconditional copula regression curve for $C^{(2)}$. Thus, the proposed model gives regression curves that vary between clusters, and their expectation is equal to the marginal copula regression based on $C^{(2)}$.

We now suppose that the copula $C^{(3)}_{1,1:n}$ is 1-exchangeable and normal with correlation ρ_3. Then the cdf of random vector $\left\{\Phi^{-1}(W_1), \dots, \Phi^{-1}(W_n)\right\}$ is a multivariate normal distribution with an exchangeable correlation matrix whose entries are 1 on the diagonal and ρ_3 off the diagonal, where $W_j = C^{(2)}_{2|1}\{G(Y_j)|F(x_j)\}, j \in \{1, \dots, n\}$. Using standard properties of the multivariate normal distribution, the conditional distribution of $\Phi^{-1}(W_n)$, knowing $\{\Phi^{-1}(W_j), \dots, \Phi^{-1}(W_{n-1})\}^{\mathrm{T}}$, is univariate normal with mean μ_0 and variance σ_0^2, defined by

$$\mu_0 = \frac{(n-1)\rho_3\sum_{j=1}^{n-1}\Phi^{-1}(W_j)/(n-1)}{1+(n-2)\rho_3}, \quad \sigma_0^2 = \frac{(1-\rho_3)\{1+(n-1)\rho_3\}}{1+(n-2)\rho_3}, \tag{13}$$

as derived in Rivest et al. (2016). In other words $\Pr\{\Phi^{-1}(W_n) \le z|x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n\} = \Phi\{(z-\mu_0)/\sigma_0\}$ for $z \in \mathbb{R}$. As $Y_n = G^{-1}[\{C^{(2)}_{2|1}\}^{-1}\{W_n|F(x_n)\}]$ is an increasing function of W_n, simple transformations lead to the following expressions:

- The conditional distribution of Y_n,

$$G_P(y) = \Pr\{Y_n \le y|x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n\} = \Phi\left(\frac{\Phi^{-1}[C^{(2)}_{2|1}\{G(y)|F(x_n)\}] - \mu_0}{\sigma_0}\right), \ y \in \mathbb{R}.$$

- The conditional density, obtained by differentiating $G_P(y)$ with respect to y,

$$g_P(y) = g(y)\frac{c^{(2)}\{F(x_n), G(y)\}}{\sigma_0} \times \exp\left\{-\frac{1}{2\sigma_0^2}\left(\Phi^{-1}\left[C^{(2)}_{2|1}\{G(y)|F(x_n)\}\right] - \mu_0\right)^2 + \frac{1}{2}\Phi^{-1}\left[C^{(2)}_{2|1}\{G(y)|F(x_n)\}\right]^2\right\},\ y \in \mathbb{R}. \tag{14}$$

- The quantile function, the inverse of $G_P(y)$,

$$G_P^{-1}(w) = G^{-1}\left(\{C^{(2)}_{2|1}\}^{-1}[\Phi\{\mu_0 + \sigma_0\Phi^{-1}(w)\}|F(x_n)]\right),\ w \in (0,1).$$

The predictor (12) when $C^{(3)}_{1,1:n}$ is a normal copula is given by the integral of the quantile function

$$\mathrm{E}_P(Y_n) = \int_{\mathbb{R}} G^{-1}\left[\{C^{(2)}_{2|1}\}^{-1}\{\Phi(\mu_0 + \sigma_0 z)|F(x_n)\}\right]\phi(z)dz, \tag{15}$$

where $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$. This conditional expectation is easily evaluated using the Gauss–Hermite quadrature method (Stefanski and Boos, 2002). When $\mu_0 = 0$ and $\sigma_0^2 = 1$, Equation (15) gives the marginal copula regression prediction for $C^{(2)}$. The variance σ_0^2 decreases with n; it goes from $1 - \rho_3^2$ at $n = 2$ to $1 - \rho_3$ as $n \to \infty$. In many instances, ρ_3 is small and the regression curve at $\mu_0 \approx 0$ is nearly equal to the marginal copula regression for $C^{(2)}$.

These formulas contain, as special cases, standard results for the normal linear mixed model with random intercepts $ML1$, defined by Equation (9). In this situation, the copula $C^{(2)}$ and the marginal distributions for X and Y are normal. It is then possible to show that $\Phi^{-1}(w_j) = r_j/\sigma_r$, where $r_j = y_j - \beta_0 - \beta_1 x_j$ is the fixed effect residual for data point j and $(\beta_0, \beta_1, \sigma_r)$ are defined in (9) and (10). Thus, $\mu_0 = (n-1)\rho_3\bar{r}/[\sigma_r\{1 + (n-2)\rho_3\}]$, where $\bar{r}$ is the average residual. The conditional distribution of Y_n, $G_P(y)$, is then normal. Its expectation is the BLUP (best linear unbiased predictor) for Y_n, $\beta_0 + \beta_1 x_n + (n-1)\rho_3\bar{r}/\{1 + (n-2)\rho_3\}$ (Rao and Molina, 2015, p. 175). Its variance is $\sigma_r^2 \times \sigma_0^2 = \sigma_e^2[1 + \rho_3/\{1 + (n-2)\rho_3\}]$, where σ_0^2 and σ_e^2 are, respectively, defined in (13) and (10).

When n is large, $\mu_0 \approx \sum_{j=1}^{n-1}\Phi^{-1}(W_j)/(n-1)$ varies between clusters according to a $\mathcal{N}(0, \rho_3)$ distribution. This defines the range of possible cluster-specific regression curves. Using Equation (15), we construct the prediction curves for three copulas $C^{(2)}$, viz. the Gumbel, Frank and Clayton, when the margins F and G are the standard normal. The curves are presented in Figure 3. They correspond to Kendall's tau of $(0.3, 0.6, 0.9)$ for each copula $C^{(2)}$. The dependence parameters are deduced from Kendall's tau following (Joe, 2014, pp. 58, 166, 168) and using the function `iTau` of the R-package **copula** (Kojadinovic and Yan, 2010). We also set the cluster size to $n = 21$ and the Kendall's tau of $C^{(3)}_{1,1:n}$ to 0.1 corresponding to $\rho_3 = 0.16$; such a weak residual dependency is found in many applications, such as in the analysis presented in Section 6. Prediction curves are plotted for the quantiles $q \in \{1/10, 5/10, 9/10\}$ of $\sum_{j=1}^{n-1}\Phi^{-1}(W_j)/(n-1)$. When $n = 21$ and $\rho_3 = 0.16$, the values of (μ_0, σ_0^2) for the three curves in each figure are $(-0.41, 0.87)$, $(0, 0.87)$ and $(0.41, 0.87)$, respectively. In each plot, the middle curve, in green, is approximately equal to the marginal copula regression curve.

The graphs exemplify the impact of the dependency in $C^{(2)}$ on the prediction curves. Indeed, when the copula $C^{(2)}$ is normal, we obtain parallel regression lines. For Clayton's copula, the regression curves are curved and close together for small values of x. When the $C^{(2)}$ dependencies increase, the regression curves tend to unique straight lines. This agrees with the conditional form of the model given in (9) whose random part tends to 0 as ρ_{U_1,V_1} increases to 1.

4 Copula selection and parameter estimation for a 2-exchangeable model

The dataset to analyze is $\{(X_{ij}, Y_{ij})^{\mathrm{T}} : i = 1, \ldots, m;\ j = 1, \ldots, n_i\}$, where the index i is for clusters and the index j is for units within clusters. The clusters are assumed to be independent, and the joint density of the $2n_i$ variables in cluster i is as given in Proposition 3. It is determined by two marginal distributions $F(x|\alpha)$, $G(y|\beta)$ and by a copula density given in Equation (11), which involves a bivariate copula density, $c^{(2)}(u, v; \delta_2)$, for the marginal

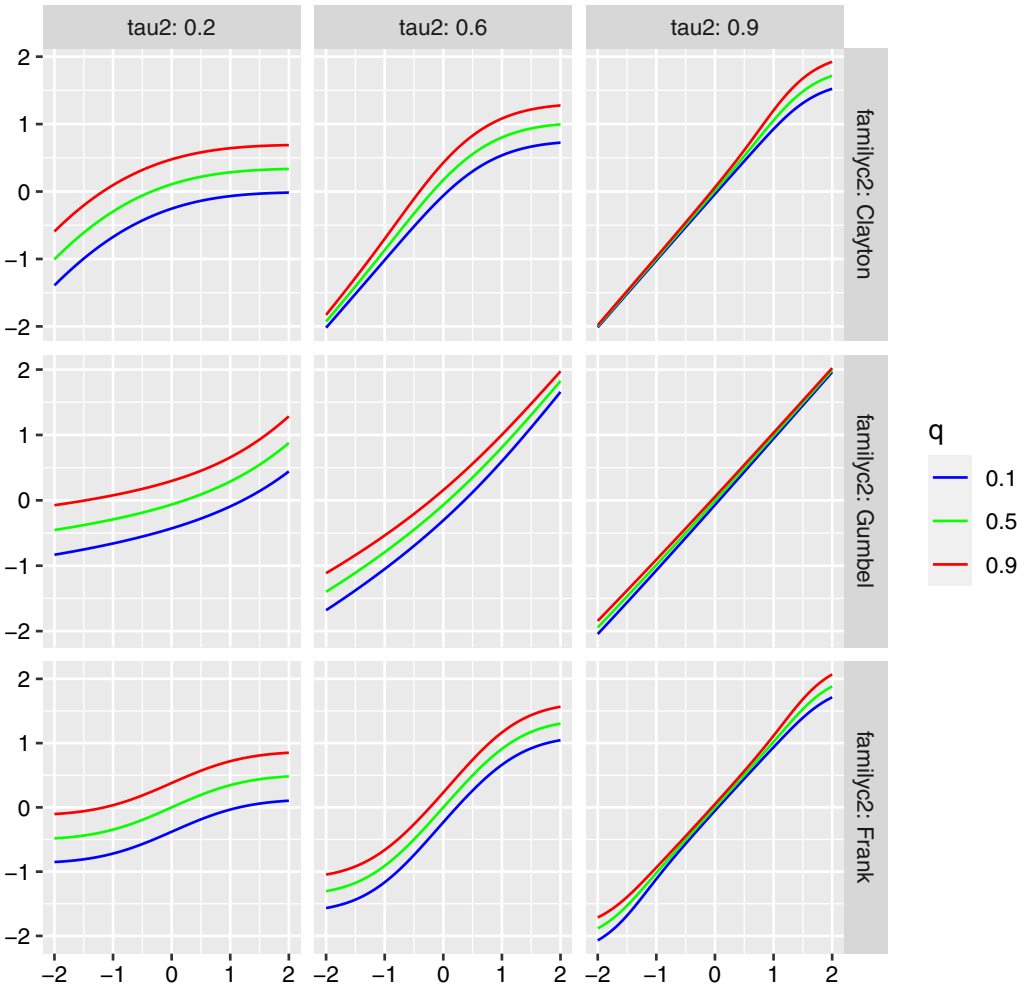

Figure 3: Prediction curves $E_P(Y_n)$ for three copulas $C^{(2)}$ when the margins F and G are standardized normal distributions.

relationship between X and Y and two 1-exchangeable copula families with densities $c^{(1)}_{1,1:n}(u_1,\dots,u_n;\delta_1)$ and $c^{(3)}_{1,1:n}(w_1,\dots,w_n;\delta_3)$. The goal of this section is to identify parametric families for the five components of this model and to estimate their parameters that are put in vector $\theta=(\alpha,\beta,\delta_1,\delta_2,\delta_3)$. Following the model-building procedure presented in Joe (2014, chap. 5) and Czado (2019, chap. 7 and 8), the model components are selected sequentially, using techniques that are presented in this section.

4.1 Determination of the model components

The determination of the marginal distributions is done independently for the two margins. Competing models are compared on the basis of the Akaike information criterion (AIC) calculated as if the units were independent. Preliminary, or IFM, estimators (Joe, 2014, section 5.5) $(\tilde{\alpha},\tilde{\beta})$ are obtained by maximizing

$$\mathcal{L}_F=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\{f(x_{ij}|\alpha)\}\text{ and }\mathcal{L}_G=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\{g(y_{ij}|\beta)\}.$$

The next step is to calculate the pseudo-observations, defined by

$$\tilde{u}_{ij}=F(x_{ij}|\tilde{\alpha}),\qquad \tilde{v}_{ij}=G(y_{ij}|\tilde{\beta})\,. \tag{16}$$

The $N=n_1+\dots+n_m$ pairs $(\tilde{u}_{ij},\tilde{v}_{ij})$ are then used to select a copula $C^{(2)}$ for the (X,Y) dependency. The within-cluster dependency impacts the variances of the estimators; however, this is overlooked at this stage, and

standard methods, proposed in Joe (2014, chap.5) and Czado (2019, chap. 7 and 8), are used to select a copula family $C^{(2)}$ for the pooled bivariate sample. Competing models are compared on the basis of their AIC, and an IFM estimator $\tilde{\delta}_2$ of the parameter of the selected copula family is obtained by maximizing

$$\mathcal{L}_2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left[c^{(2)}(\tilde{u}_{ij}, \tilde{v}_{ij}; \delta_2)\right].$$

To assess the within-cluster dependency associated with copula families $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$, we use the unit-level version of the exchangeable Kendall's tau introduced in Romdhani et al. (2014). It is evaluated using the proportion of concordant pairs $\{(x_{ij}, x_{i\ell}), (x_{kr}, x_{ks})\}$ among the $\sum_{i>k} n_i(n_i-1)n_k(n_k-1)$ possible pairs of ordered observations coming from different clusters. Models can also be compared on the basis of their AIC, and an IFM estimator of δ_1 is obtained by maximizing

$$\mathcal{L}_1 = \sum_{i=1}^{m} \log\left\{c^{(1)}_{1,1:n_i}(\tilde{u}_{i1}, \ldots, \tilde{u}_{in_i}; \delta_1)\right\}.$$

The selection of the copula family $C^{(3)}_{1,1:n}$ is based on the pseudo-observations $\tilde{w}_{ij} = C^{(2)}_{2|1}(\tilde{v}_{ij}|\tilde{u}_{ij}; \tilde{\delta}_2)$. The exchangeable Kendall's tau evaluated on $\{\tilde{w}_{ij} : i = 1, \ldots, m; j = 1, \ldots, n_i\}$ can be used to assess the within-cluster dependency of the residuals. An IFM estimator $\tilde{\delta}_3$ is obtained by maximizing

$$\mathcal{L}_3 = \sum_{i=1}^{m} \log\left\{c^{(3)}_{1,1:n_i}(\tilde{w}_{i1}, \ldots, \tilde{w}_{in_i}; \delta_3)\right\}.$$

As noted in Joe (2014, section 5.5), the five estimating equations associated with the IFM estimator $\tilde{\theta}$ can be combined into a single multivariate estimating equation. The standard asymptotic theory, as the number of clusters m goes to ∞, presented in Tsiatis (2006, p. 30), applies as long as the cluster sizes n_i are bounded. It shows that the limiting asymptotic distribution of $\sqrt{m}(\tilde{\theta} - \theta)$ is a centred multivariate normal with a sandwich covariance matrix that accounts for the within-cluster dependence.

4.2 Maximum likelihood (ML) estimation of the parameters

Once the parametric families for the five model components have been identified, optimal estimators of the parameters are obtained by ML. This section discusses the properties of the ML estimator for the parameter vector θ.

The log-likelihood for θ is equal to

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\{f(x_{ij}|\alpha)\} + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left[c^{(2)}\{F(x_{ij}|\alpha), G(y_{ij}|\beta); \delta_2\}\right] \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\{g(y_{ij}|\beta)\} + \sum_{i=1}^{m} \log\left[c^{(1)}_{1,1:n_i}\{F(x_{i1}|\alpha), \ldots, F(x_{in_i}|\alpha); \delta_1\}\right] \\ &+ \sum_{i=1}^{m} \log\left[c^{(3)}_{1,1:n_i}\left[C_{2|1}\{G(y_{i1}|\beta)|F(x_{i1}|\alpha)\}, \ldots, C_{2|1}\{G(y_{in_i}|\beta)|F(x_{in_i}|\alpha)\}; \delta_3\right]\right].\end{aligned}$$

This function is easily maximized once parametric families for the two margins in the model and the three copulas are selected. This yields $\hat{\theta}$, the ML estimator of the parameter vector. This maximization is carried out using an optimizer such as the R-function `optim`. The negative Hessian of $\mathcal{L}(\theta)$, evaluated at $\hat{\theta}$, is the observed Fisher information for the model. Its inverse is the asymptotic covariance matrix of $\hat{\theta}$. It can be used to calculate standard errors for all the parameters that have been estimated. At this stage, likelihood-ratio tests comparing nested candidate parametric families for $C^{(2)}$ can also be carried out to validate the IFM model-selection step that ignored the within-cluster dependency.

When the margins $F(x)$, $G(y)$ and the copula $C^{(2)}$ are normal, the likelihood can be split into a marginal likelihood for the parameters of $F(x)$ and $C^{(1)}_{1,1:n}$ times a conditional likelihood for the other parameters. The standard normal mixed linear model falls into that category, as its parameters are estimated using a conditional

likelihood. In general, the parameters are intertwined in a complicated way, and their estimation relies on the log-likelihood $\mathcal{L}(\boldsymbol{\theta})$.

5 Monte Carlo investigations of the 2-exchangeable model

This section compares the sampling properties of the IFM and ML estimators of the parameters of a 2-exchangeable model. Then the IFM selection of the copula $C^{(2)}$ is investigated. Finally, out-of-sample predictions for the proposed model are compared with those of competing models such as normal linear mixed models and a simple copula regression, without cluster effects. Throughout this section, the margins F and G are normal distributions with mean $\mu_1 = \mu_2 = 0$ and variance $\sigma_1^2 = \sigma_2^2 = 1$. The 1-exchangeable copulas $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$ are normal with respective correlations $\rho_1 = 0.31$ and $\rho_3 = 0.16$. Correlations are parameterized in terms of their logit, $\eta = \log\{\rho/(1-\rho)\}$, to avoid bounded parameter spaces. The simulations use $B = 1000$ repetitions.

5.1 Sampling properties of IFM and ML estimators of a 2-exchangeable model

For a 2-exchangeable model with parameters $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are the ML and the IFM estimators. This section investigates their sampling properties for finite m using Monte Carlo simulations. The expectation and the variance of an estimator $\hat{\psi}$, an entry of either $\hat{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}$, are approximated by

$$\mathbb{E}_B(\hat{\psi}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_b, \qquad \mathbb{V}_B(\hat{\psi}) = \frac{1}{B-1}\sum_{i=1}^{B}\{\hat{\psi}_b - \mathbb{E}_B(\hat{\psi})\}^2.$$

Three copulas $C^{(2)}$ are investigated: the normal copula, the Clayton copula and a Khoudraji copula, introduced in Genest et al. (1998), that features an asymmetric relationship between U and V. It is defined by

$$C^{(2)}(u, v; \rho, \kappa) = u^{1-\kappa}C_\rho(u^\kappa, v), \quad u, v \in [0,1], \tag{17}$$

where C_ρ is a normal copula with correlation $\rho \in (0,1)$, and $\kappa \in (0,1)$ is the asymmetry parameter. In the maximization program, κ is parameterized in terms of its logit, η_κ. Two values for the Kendall's tau τ_2 of copula $C^{(2)}$, 0.4 and 0.6, are considered. Two sample sizes, $m = 10, 50$, are investigated; their corresponding cluster sizes are $n = 30$ and $n = 18$, respectively.

Detailed simulation results, including simulations with unequal cluster sizes, are presented in the Supplementary Material. Table 1 is representative of the general findings. All estimators have negligible biases, so the discussion focuses on variances. As expected, the strength of the U-V association in $C^{(2)}$ does not impact the precision of the two estimators for η_1. The loss of precision for the IFM estimators is larger for the parameters of copula $C^{(2)}$ than for the parameters of the other two copula families. This loss is more important at $m = 10$ than at $m = 50$. Unequal sample sizes are associated with a small loss of precision for all estimators, especially at $m = 10$. This loss is, in general, more important for IFM estimators than for ML estimators. Additional results on the sampling properties of ML and IFM estimators can be found in Akpo (2022).

Table 1: Expectations of the estimators with their variances multiplied by 10 in parentheses, when $C^{(2)}$ is the Khoudraji copula.

τ_2	m	Method	$\eta_1 = -0.80$	η_ρ	η_κ	$\eta_3 = -1.69$
0.4	10	ML	−0.90 (1.83)	0.80 (0.55)	1.51 (2.42)	−1.85 (2.13)
		IFM	−0.95 (2.08)	0.75 (0.65)	1.49 (3.25)	−1.89 (2.49)
	50	ML	−0.82 (0.51)	0.77 (0.12)	1.51 (0.98)	−1.73 (0.63)
		IFM	−0.83 (0.58)	0.76 (0.18)	1.51 (1.44)	−1.74 (0.79)
0.6	10	ML	−0.86 (1.77)	1.49 (0.36)	3.39 (5.30)	−1.81 (2.20)
		IFM	−0.94 (2.09)	1.45 (0.38)	3.62 (6.34)	−1.88 (2.74)
	50	ML	−0.81 (0.49)	1.47 (0.11)	3.48 (2.14)	−1.77 (0.73)
		IFM	−0.83 (0.63)	1.47 (0.14)	3.50 (2.79)	−1.79 (0.86)

Table 2: Percentage of simulations when a copula $C^{(2)}$ is selected as a function of the true copula when $m = 10$, $n = 30$, for $\tau_2 = 0.4$ with values for $\tau_2 = 0.6$ in parentheses.

		Selected $C^{(2)}$				
		Normal	Clayton	Gumbel	Frank	Khoudraji
True $C^{(2)}$	Normal	81.2 (95.0)	1.0 (0.0)	4.8 (0.6)	7.8 (0.0)	5.2 (4.4)
	Clayton	0.6 (0.0)	99.2 (100)	0.0 (0.0)	0.0 (0.0)	0.2 (0.0)
	Gumbel	5.4 (0.6)	0.0 (0.0)	89.2 (98.2)	2.4 (0.0)	3.0 (1.2)
	Frank	8.4 (0.0)	0.8 (0.0)	2.4 (0.0)	86.4 (100)	2.4 (0.0)
	Khoudraji	26.6 (12.6)	0.0 (0.0)	5.2 (1.2)	2.6 (1.0)	65.6 (85.2)

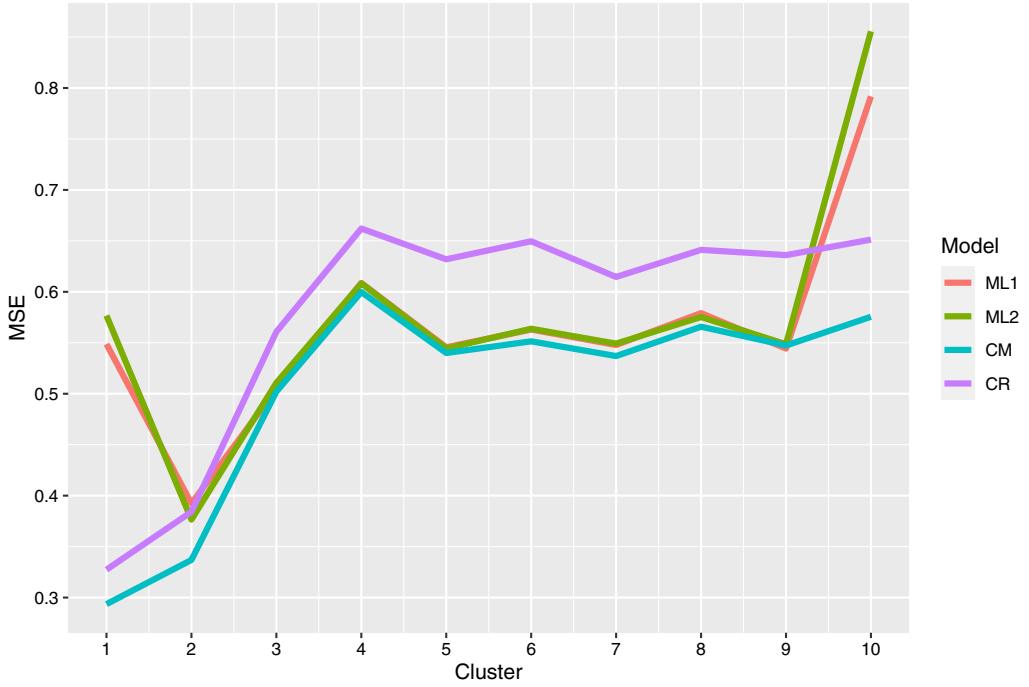

Figure 4: Mean squared prediction errors for 10 out-of-sample y values obtained with four models.

5.2 A simulation study of the IFM selection of copula $C^{(2)}$

We now investigate the ability of the IFM method to select the copula $C^{(2)}$. The simulations use $m = 10$ clusters of size $n = 30$; simulations for $m = 50$ and $n = 18$ are presented in the Supplementary Material. Five candidate copulas $C^{(2)}$ are considered: the normal, the Clayton, the Frank, the Gumbel and the Khoudraji copula, specified in Section 5.1. Samples are simulated using the five possible $C^{(2)}$ copulas, with $\tau_2 = 0.4, 0.6$. The five copulas are fitted to the pseudo-observations, defined in (16), of each simulated sample. The IFM-selected copula is the one with the smallest AIC. The results are reported in Table 2. Overall, the true copula $C^{(2)}$ is selected for most simulated samples, and the percentage of correct selection increases with τ_2. This supports the proposal of Section 4 to use the IFM method to select the components of the 2-exchangeable copula model.

5.3 Comparison of out-of-sample predictions: the 2-exchangeable model vs competing alternatives

This section evaluates the predictive ability of the proposed model in a conditional simulation where the x-values are fixed. It uses the $m = 10,\ n = 30$ scenario, with $C^{(2)}$ equal to the Clayton copula with $\tau_2 = 0.4$. One data point in each cluster is put in an out-of-sample validation set. The x-values in the validation set are -2.00, -1.56, -0.76, -0.46, 0.24, 0.35, 0.00, 1.14, 0.91 and 1.73 for clusters 1 to 10. An in-sample simulated dataset has $m = 10,\ n = 29$. Four models are fitted to this dataset: a normal linear mixed model with random intercepts ($ML1$); a linear mixed model with random slopes and intercepts ($ML2$); the proposed 2-exchangeable

normal-Clayton-normal model (CM); and a simple Clayton copula model (CR), without cluster effects. Predictions for the 10 out-of-sample y-values are calculated for the four models. For CM, a prediction is calculated using Equation (15), while Equation (12), with $c^{(3)}_{1,1:n} \equiv 1$, is used for CR. For the normal linear mixed models, the function `predict` of the R-package **nlme** of Pinheiro et al. (2012) is used.

The mean squared error (MSE) of prediction $\hat{y}_k$ is, for $k \in \{1, \ldots, 10\}$, evaluated as

$$MSE(y_k) = \frac{1}{B}\sum_{b=1}^{B}(y_{kb} - \hat{y}_{kb})^2,$$

where y_{kb} is the simulated value for y for the kth out-of-sample point and $\hat{y}_{kb}$ is a predicted value obtained using one of the four models investigated. The results are presented in Figure 4. As expected, CM predictions have the smallest MSEs. The results for $ML1$ and $ML2$ are similar. Their predictions MSEs and those for CM are similar for the middle x-values. Considering Figure 3, Clayton regression curves are not linear; the $ML1$ and $ML2$ predictions for the more extreme out-of-sample x-values, -2 and 1.73 in clusters 1 and 10, have large biases and large MSEs. The CR predictions, on the other hand, do not use cluster effects; their MSEs are slightly larger than those for CM.

6 Modelling math grades with a 2-exchangeable copula model

This section revisits a dataset discussed in Goldstein (2011, chap. 2), available in the R-package **lmeresampler** (Loy et al., 2023) under `data(jsp728)`. It consists of math grades in the fourth and seventh years measured on $N = n_1 + \ldots + n_{48} = 728$ students in $m = 48$ primary schools. The cluster sample sizes n_i vary between 4 and 40. The fourth-year mark (X) and the seventh-year mark (Y) vary between 0 and 40. We map them to the (0,1) interval using the transform $M \to (M + 1/2)/41$. To break ties in the grades, a small random perturbation was added to each one; this is useful for the computation of the exchangeable Kendall's tau.

One objective of this section is to construct predictive models for Y given X whose support is $(0, 1)$. Another objective is to investigate whether the 2-exchangeable copula model can capture the nonlinearity seen in Figure 5, which gives a scatter plot of the $N = 728$ data points and a nonparametric regression curve.

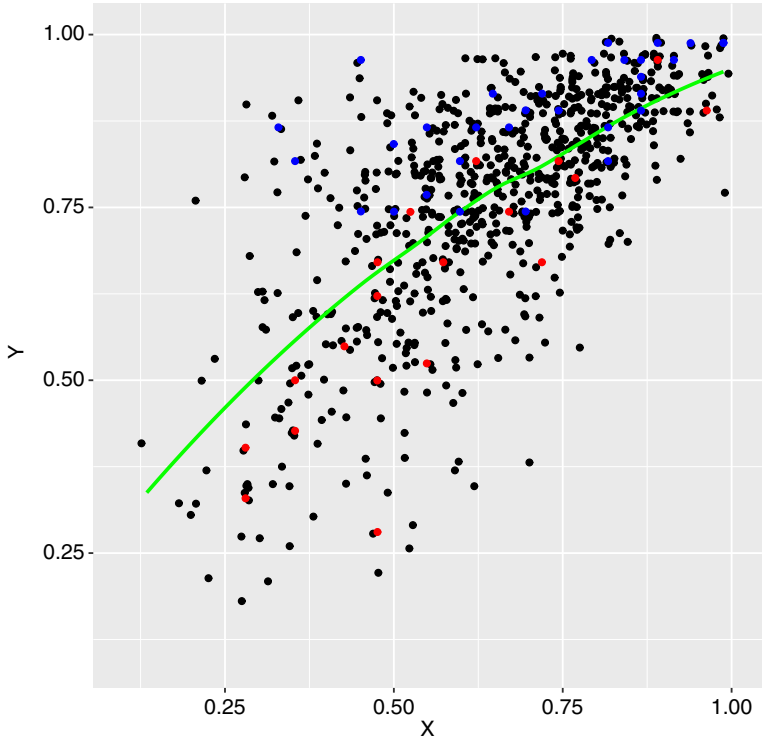

Figure 5: Scatter plot of Y (seventh-year mark) versus X (fourth-year mark) with a nonparametric regression curve. Schools 1 and 30 are identified with, respectively, red and blue points.

Table 3: Fit of the $\mathcal{B}$ and of the $\mathcal{GB}3$ distributions to the two margins.

Variable	Model	$\tilde{\theta}$	*pse*	AIC
X	$\mathcal{B}$	(4.280,2.332)	(0.222,0.115)	−542.88
	$\mathcal{GB}3$	(4.280,2.330,1)	(0.222,0.115,NA)	−540.80
Y	$\mathcal{B}$	5.24,1.79	(0.28,0.08)	−789.57
	$\mathcal{GB}3$	(2.616,2.319,0.29)	(0.303,0.241,0.069)	−826.96

The goal is to contrast an analysis carried out with copulas with the one reported in Goldstein (2011, chap. 2) which is based on standard normal linear mixed models.

6.1 Selection of the marginal distributions for X and Y

The candidate families for F and G are the beta (denoted $\mathcal{B}$) and the generalized beta (denoted $\mathcal{GB}3$) distributions (Cockriel and McDonald, 2018). If X has a $\mathcal{B}(\alpha,\beta)$ distribution, then $Y = X/\{\lambda + (1-\lambda)X\}$ has for $\lambda \in (0,1)$ a $\mathcal{GB}3(\alpha,\beta,\lambda)$ distribution whose density is given by

$$\frac{\lambda^{\alpha}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{y^{\alpha-1}(1-y)^{\beta-1}}{\{1-(1-\lambda)y\}^{\alpha+\beta}}, \qquad 0 \le y \le 1\cdot$$

Table 3 compares the fit of these two distributions to the two margins. As stated in Section 4.1, this preliminary analysis does not account for the within-cluster dependency. Thus, in Table 3 *pse*, for pseudo-standard error, ignores the within-school dependency.

The best fitting models are, respectively, the beta and the generalized beta for X and Y.

6.2 Selection of the bivariate copula $C^{(2)}$

The first step is to calculate the pseudo-observations $[\tilde{u}_{ij}, \tilde{v}_{ij} : j \in \{1,\dots,n_i\}, i \in \{1,\dots,m\}]$, given by (16). Kendall's tau is 0.49 ($pse = 0.03$), so there is a relatively strong association between the two variables. Following Joe (2014, chap. 1), Kendall's tau is calculated for the subsamples in the four quadrants of the unit square. This reveals a stronger association for large grades than for smaller ones. Also, a 0.1 difference between Kendall's tau for the upper left and lower right quadrants suggests that some of the asymmetry seen in Figure 5 remains once the margins' effect has been factored out.

Several copulas were fitted to this bivariate sample using the function `BiCopEst` in the R package **VineCopula** (Nagler et al., 2021) and the function `fitCopula` in **copula**. To capture the asymmetry in the data, we used the Khoudraji device, defined in (17), to create asymmetric alternatives. The best fitting copula in Table 4 is the survival Khoudraji normal copula given by

$$C^{(2)}(u,v|\rho_2,\kappa_1,\kappa_2) = u + v - 1 + (1-u)^{1-\kappa_1}(1-v)^{1-\kappa_2}C_{\rho_2}\{(1-u)^{\kappa_1},(1-v)^{\kappa_2}\},$$

where C_{ρ_2} is the bivariate normal copula with correlation ρ_2. The density of a copula in this three-parameter family can be evaluated using the functions `KhoudrajiCopula` and `pCopula` of the package **copula**. Its conditional distribution, $w = \partial C^{(2)}(u,v|\rho_2,\kappa_1,\kappa_2)/\partial u$, has the following explicit form:

$$\begin{aligned} w = {} & 1 - (1-\kappa_1)(1-u)^{-\kappa_1}(1-v)^{1-\kappa_2}C_{\rho_2}\{(1-u)^{\kappa_1},(1-v)^{\kappa_2}\} \\ & - \kappa_1(1-v)^{1-\kappa_2}\Phi\left[\frac{\Phi^{-1}\{(1-v)^{\kappa_2}\} - \rho_2\Phi^{-1}\{(1-u)^{\kappa_1}\}}{\sqrt{1-\rho_2^2}}\right]. \end{aligned} \tag{18}$$

In Table 4, Survival Khoudraji-Normal2 refers to a two-parameter version of this copula obtained by setting $\kappa_2 = 1$.

Table 4: Paramater estimates, their pseudo-standard errors *pse* and the AIC for several copulas for the (u, v) relationship.

Copula $C^{(2)}$	$\tilde{\theta}$	*pse*	AIC
Normal	0.682	0.016	−454.32
Survival-Gumbel	1.892	0.057	−459.5
Survival Khoudraji-Normal	(0.791, 0.822, 0.960)	(0.023, 0.043, 0.029)	−474.10
Survival Khoudraji-Normal2	(0.762, 0.837)	(0.021, 0.045)	−468.64

Table 5: Parameter estimates, *pse*s and the AIC for three candidates for $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$.

	$C^{(1)}_{1,1:n}$			$C^{(3)}_{1,1:n}$		
Family	$\tilde{\theta}$	*pse*	AIC	$\tilde{\theta}$	*pse*	AIC
Frank	0.251	0.138	−5.69	0.935	0.179	−69.76
Gumbel	1.040	0.022	−4.64	1.109	0.024	−66.17
Normal	0.064	0.025	−12.76	0.167	0.035	−80.03

Table 6: Maximum likelihood estimates of the parameters for the full model.

Component	$\hat{\theta}$	*se*
F ($\mathcal{B}$)	$(\hat{\alpha}_1, \hat{\beta}_1) = (4.271, 2.359)$	(0.235, 0.124)
G ($\mathcal{GB}3$)	$(\hat{\alpha}_2, \hat{\beta}_2, \hat{\lambda}) = (2.457, 2.470, 0.248)$	(0.245, 0.254, 0.052)
$C^{(1)}_{1,1:n}$ (Normal)	$\hat{\rho}_1 = 0.063$	0.026
$C^{(2)}$ (Survival Khoudraji-Normal)	$(\hat{\rho}, \hat{\kappa}_1, \hat{\kappa}_2) = (0.795, 0.822, 0.959)$	(0.024, 0.046, 0.029)
$C^{(3)}_{1,1:n}$ (Normal)	$\hat{\rho}_3 = 0.161$	0.040

6.3 Selection of the 1-exchangeable copula families $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$

The exchangeable Kendall's tau for X and Y are, respectively, 0.046 ($se = .017$) and 0.095 ($se = 0.035$), showing a stronger schools effect in the seventh year. The fits of several families of copulas for the joint distribution of $\{\tilde{u}_{ij}\}$ and $\{\tilde{w}_{ij}\}$, evaluated using Equation (18), are compared by maximizing the pseudo-log-likelihoods $\mathcal{L}_1$ and $\mathcal{L}_3$. The results are reported in Table 5. The normal family is the best choice for both $C^{(1)}_{1,1:n}$ and $C^{(3)}_{1,1:n}$.

To complete the analysis, the MLEs of the 10 parameters of the proposed model are evaluated. The IFM parameter estimate for κ_2 is very close to 1. We first carry out a likelihood ratio test for $H_0 : \kappa_2 = 1$ using the full likelihood. This gives $\chi^2_{1,obs} = 8.2$ for a P-value of 0.4%. Thus, the full model, with 10 parameters, is definitive; the MLEs and their standard errors (*se*) of the parameters are reported in Table 6.

The fit of the final model, summarized in Table 6, is illustrated using two schools, number 1, with $n_1 = 19$, and number 30, with $n_{30} = 31$. Their respective values of μ_0, defined in (13), are −0.464 and 0.810; this means that, for the same fourth-year mark, the seventh-year grades in school 30 tend to be higher than in school 1 as seen in Figure 5. This can be seen in Figure 6, which gives the two regression curves constructed using Equation (15).

Conditional residuals for the fitted copula models can be defined as $y_{ij} - \hat{y}_{ij}$, where $\hat{y}_{ij}$ is evaluated as the predicted value at x_{ij} through Equation (15), where the values of μ_0 and σ_0, given in (13), are those for school i. The conditional residuals of the copula model *CM* can be compared with those of mixed linear models, *ML*1 and *ML*2, and of the simple copula regression *CR* considered in Section 5.3. The MSEs and inter quartile ranges (IQRs) of the residuals for *CM*, *ML*1 and *ML*2, *CR* are (0.0107, 0.0112, 0.0101, 0.0137) and (0.108, 0.1186, 0.1101, 0.1289), respectively. Thus, in agreement with the IFM selection of the model components, the fits of *ML*1 and *CR* are poor. The fits of *CM* and *ML*2 are very similar, as the latter captures the between-school change in slope that can be seen in Figure 6. *CM* has a larger residual MSE; however, it has a smaller residual IQR and it gives an absolute residual smaller than that of *ML*2 for 52% of the data points.

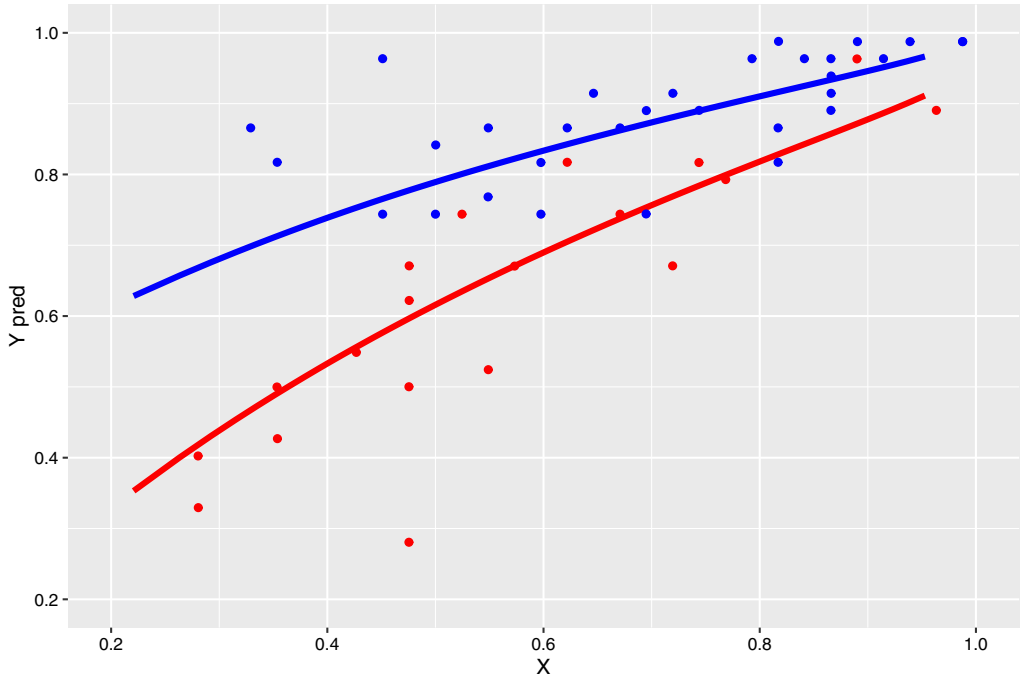

Figure 6: Scatter plots and regression curves for school 1 (in red) and school 30 (in blue).

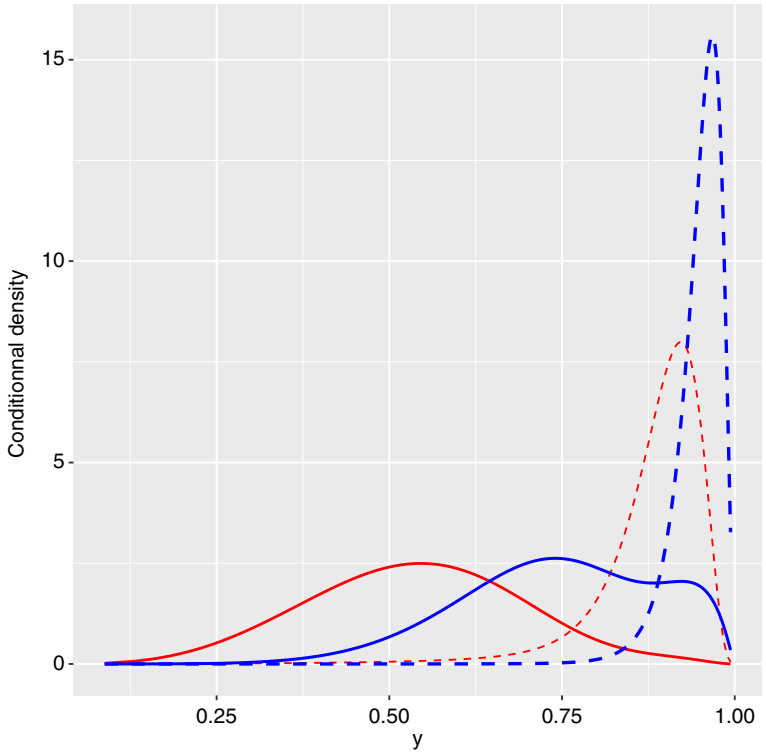

Figure 7: Densities for the predicted Y values for $X = 0.4$ (continuous lines) and $X = 0.9$ (dashed lines) for schools 1, in red, and 30, in blue.

Table 7: 95% prediction intervals for $X = 0.4$ and $X = 0.9$ in schools 1 and 30.

Model	CM		$ML1$	
School	$X = 0.4$	$X = 0.9$	$X = 0.4$	$X = 0.9$
1	(0.23, 0.89)	(0.66,0.97)	(0.33, 0.76)	(0.65,1.08)
30	(0.44,0.97)	(0.86,0.99)	(0.46,0.89)	(0.78,1.21)

Note also that $ML2$ has convergence problems; the likelihood is maximized when the correlation between the random intercepts and slopes is -1.

The key advantage of the 2-exchangeable copula model is that it allows prediction intervals for Y to depend on both the school and the X-value. This is illustrated in Figure 7, which gives the predictive densities for mark Y, given by formula (14), for $X = 0.4$ and $X = 0.9$ in schools 1 and 30. The variability of Y is larger at $X = 0.4$ and in school 1. As discussed in Section 3.3, the predictive density for $ML1$ is always normal and its variance does not vary much between clusters. Prediction intervals for Y in schools 1 and 30 are given in Table 7, which compares CM and $ML1$.

All $ML1$ prediction intervals have the same length. Two of them go beyond the maximum grade of 1. The CM prediction intervals, on the other hand, are more flexible. They are all within the $(0, 1)$ interval and their lengths at $x = 0.9$ are much smaller than at $x = 0.4$. This illustrates the flexibility of the proposed 2-exchangeable model.

7 Conclusion

The 2-exchangeable copula model proposed in this article provides flexible methods to predict the variable Y knowing X in a hierarchical dataset. A key feature of the proposed methodology, highlighted in Section 4, is the flexibility of the predictive densities for Y given X. Its shape and its support can depend on the known X value for Y and on the cluster. An outstanding problem is whether the proposed model can be generalized to $d - 1 > 1$ continuous explanatory variables. The key to such a generalization is the availability of flexible $(d-1)$-exchangeable families $C^{(1)}_{d-1,1:n}$ for the joint distribution of the explanatory variables for all the units in a cluster. An elliptical copula, with a correlation matrix given in (3), could be used. This specifies partially the copula $C^{(2)}$ for the joint distribution of the X and the Y variables on a unit. These problems will be the subject of future investigations.

Data sharing

The dataset analyzed in Section 6 is available in the R package lmeresampler (Loy et al., 2023) under data(jsp728). The R code for carrying out most of the analyses presented in Section 6 can be found at https://github.com/Gabatsarl/R-Program-Copula-CJS/tree/main.

Acknowledgements

The authors are grateful to Étienne Marceau for his detailed comments on a preliminary version of this article and to the referees whose comments improved it.

Funding information

This research was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada and by a Canada Research Chair in Statistical Sampling and Data Analysis.

ORCID

Talagbe Gabin Akpo: https://orcid.org/0000-0003-3957-9948
Louis-Paul Rivest: https://orcid.org/0000-0003-4351-4127

Supplementary material

The Supplementary Material contains proofs for some of the results in Sections 2–4, a representation of the R-vine for the proposed model in clusters of arbitrary size, and additional simulation results.

References

Acar, E. F., Azimaee, P., and Hoque, M. E. (2019). Predictive assessment of copula models. *The Canadian Journal of Statistics*, 47(1):8–26.

Akpo, T. G. (2022). *Régression avec copules pour des données hiérarchiques*. Ph.D. thesis, Université Laval, Québec, Canada. https://corpus.ulaval.ca/server/api/core/bitstreams/b7bbebb7-76a8-43c0-bcdb-69c4c77844d0/content.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Bernard, C. and Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, 138:104–126.

Cockriel, W. M. and McDonald, J. B. (2018). Two multivariate generalized beta families. *Communications in Statistics - Theory and Methods*, 47(23):5688–5701.

Côté, M.-P., Genest, C., and Omelka, M. (2019). Rank-based inference tools for copula regression, with property and casualty insurance applications. *Insurance: Mathematics and Economics*, 89:1–15.

Crane, G. J. and Van der Hoek, J. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics*, 50(1):53–67.

Czado, C. (2019). *Analyzing Dependent Data with Vine Copulas: A Practical Guide with R*. Lecture Notes in Statistics 222, Springer Nature Switzerland, Cham, Switzerland.

Czado, C. and Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, 9:453–477.

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

Genest, C., Ghoudi, K., and Rivest, L.-P. (1998). Discussion of "Understanding relationships using copulas, by E. Frees and E. Valdez". *North American Actuarial Journal*, 2(3):143–152.

Goldstein, H. (2011). *Multilevel Statistical Models*, 4th ed., Wiley, Chichester, UK.

Grover, K., Acar, E. F., and Torabi, M. (2020). Copula-based predictions in small area estimation. *The Canadian Journal of Statistics*, 48(4):685–711.

Hofert, M., Prasad, A., and Zhu, M. (2023). Rafternet: Probabilistic predictions in multi-response regression. *The American Statistician*, 77(4):406–416.

Joe, H. (2014). *Dependence Modeling with Copulas*, Chapman and Hall, Boca Raton, Florida.

Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20.

Kumar, P. and Shoukri, M. (2007). Copula based prediction models: an application to an aortic regurgitation study. *BMC Medical Research Methodology*, 7(21):1–9.

Loy, A., Steele, S., and Korobova, J. (2023). *lmeresampler: Bootstrap methods for nested linear mixed-effects models*. R package version 0.2.4. https://CRAN.R-project.org/package=lmeresampler.

Mai, J. F. and Scherer, M. (2012). Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications, *World Scientific*, London.

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, Linear, and Mixed Models*, John Wiley & Sons, New York.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2021). *VineCopula: Statistical inference of vine copulas*. R package version 2.4.3. https://CRAN.R-project.org/package=VineCopula.

Nelsen, R. B. (2006). *An Introduction to Copulas*, Springer, New York.

Noh, H., Ghouch, A. E., and Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association*, 108(502):676–688.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2012). Nonlinear mixed-effects models. *R package version* 3. 1–89.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, John Wiley & Sons, New York.

Rivest, L.-P., Verret, F., and Baillargeon, S. (2016). Unit level small area estimation with copulas. *The Canadian Journal of Statistics*, 44(4):397–415.

Romdhani, H., Lakhal-Chaieb, L., and Rivest, L.-P. (2014). An exchangeable Kendall's tau for clustered data. *The Canadian Journal of Statistics*, 42(3):384–403.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1): 29–38.

Su, C.-L., Nešlehová, J. G., and Wang, W. (2019). Modelling hierarchical clustered censored data with the hierarchical Kendall copula. *The Canadian Journal of Statistics*, 47(2):182–203.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Received 12 June 2023 — Accepted 13 May 2024