Multivariate distributions of correlated binary variables generated by pair-copulas

Huihui Lin and N. Rao Chaganty[*]

*Correspondence:
rchagant@odu.edu
Department of Mathematics &
Statistics, Old Dominion University,
2300 Elkhorn Avenue, Norfolk
23529, USA

Abstract

Correlated binary data are prevalent in a wide range of scientific disciplines, including healthcare and medicine. The generalized estimating equations (GEEs) and the multivariate probit (MP) model are two of the popular methods for analyzing such data. However, both methods have some significant drawbacks. The GEEs may not have an underlying likelihood and the MP model may fail to generate a multivariate binary distribution with specified marginals and bivariate correlations. In this paper, we study multivariate binary distributions that are based on D-vine pair-copula models as a superior alternative to these methods. We elucidate the construction of these binary distributions in two and three dimensions with numerical examples. For higher dimensions, we provide a method of constructing a multidimensional binary distribution with specified marginals and equicorrelated correlation matrix. We present a real-life data analysis to illustrate the application of our results.

Keywords: D-vine, Mutivariate binary distributions, Multivariate probit model, Pair-copulas

Introduction

In clinical trials and research studies in health care and medicine, the endpoint of the observed data most often consists of correlated binary observations. The generalized estimating equation (GEE), introduced by Liang and Zeger ([1986](#)), has been the common statistical tool for analyzing such data. However, this method has several drawbacks. One of the drawbacks is that it uses an ambiguously defined working correlation to model the dependence in the binary observations, which could lead to misleading conclusions (Sabo and Chaganty, [2010](#)). Another drawback it is a non-likelihood approach, in the sense, it does not have an underlying joint distribution for the correlated binary observations. Other alternatives to GEEs for the analysis of correlated binary data are Markov chains (MCs) and multivariate probit (MP) models. A contrasting study of the first order MC model and the MP model was presented by Yang and Chaganty ([2014](#)). They showed that both models are asymptotically efficient, and discussed situations where one is preferable over the other.

In recent years, due to their success in other disciplines, copulas have been used to develop likelihood-based methods as another alternative to GEEs. Some researchers have

combined copulas with MC models. Escarela et al. (2009) have used Gaussian copula to construct conditional probabilities in Markov chain models in the context of longitudinal binary data. The copula-based bivariate probit models were generalized by Winkelmann (2012) replacing the Gaussian distribution by Frank and Clayton copulas. Radice et al. (2016) introduced nonlinear regression models, where non-Gaussian copulas were used to deal with the dependence between binary responses. Smith et al. (2010) showed that longitudinal continuous data can be modeled by D-vine pair-copula, and later extended their work to the discrete case using a Bayesian framework in Smith and Khaled (2012). A Gaussian copula model for integer-valued ARMA structured time series data with or without covariates was developed by Lennon (2016). Panagiotelis et al. (2017) introduced two algorithms for optimizing vine structure and pair-copula selection for discrete regular vine copulas. One of these algorithms uses a modified Akaike information criterion and the other uses predictive scores with cross-validation.

In this paper we study multivariate binary distributions generated by D-vine pair-copula models. These models are relatively easy to implement since they use only bivariate copulas, and flexible because they allow different types of bivariate copulas to model different types of dependence in the conditional distributions. We will see that the D-vine pair-copula model has some advantages over the MP model.

The organization of this paper is as follows. We first present a lucid description of the construction of bivariate and trivariate binary distributions using bivariate Gaussian, Clayton, Frank, and Gumbel copulas in "Construction of vine pair-copula binary distributions" section. We discuss comparisons between pair-copula models and the multivariate probit (MP) model in "Comparison of pair-copula and MP models" section, together with a numerical example where the vine pair-copula model overcomes the difficulties associated with the MP model. In "Extensions to four and higher dimensions" section we discuss extensions to four and higher dimensions and present a method of constructing multivariate binary distribution with specified marginals and equicorrelated structure. In "Parameter estimation" section, we discuss parameter estimation by maximum likelihood for grouped data. "Data analysis" section contains an analysis of a real-life data. We end the paper with some discussion in "Discussion" section.

Construction of vine pair-copula binary distributions

In this section, we illustrate methods of constructing multivariate binary distributions with specified correlations using vine pair-copula methods. The major advantage of these methods is that the multivariate distribution can be constructed using only bivariate copulas. The method is computationally feasible, flexible, and can accommodate various types of dependence because the bivariate copula need not be the same for various bivariate marginal and conditional distributions. We start first with the simplest cases, bivariate and the trivariate distributions, and then show how they can be extended to higher dimensions focusing on the special correlation structures useful in longitudinal and clustered binary data analysis.

Bivariate binary distributions

Consider first the case of two binary variables. Let $Y = (Y_1, Y_2)$, where the subscripts 1 and 2 possibly may indicate two sequential time points. Assume that $E(Y_i) = p_i$ for $i = 1, 2$. According to the theorem of Sklar (1959), the joint CDF of Y using a copula function

Table 1 Copula families of distribution functions

Name	Copula function	Parameter range
Gaussian	$C_1(u_1, u_2; \gamma) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \gamma)$	$-1 \leq \gamma \leq 1$
Clayton	$C_2(u_1, u_2; \alpha) = \max((u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-\frac{1}{\alpha}}, 0)$	$\alpha \in [-1, \infty) \setminus \{0\}$
Frank	$C_3(u_1, u_2; \alpha) = -\frac{1}{\alpha} \log(1 + \frac{(e^{-\alpha u_1} - 1)(e^{-\alpha u_2} - 1)}{e^{-\alpha} - 1})$	$\alpha \neq 0$
Gumbel	$C_4(u_1, u_2; \alpha) = e^{-[(-\log u_1)^\alpha + (-\log u_2)^\alpha]^{\frac{1}{\alpha}}}$	$\alpha \geq 1$
Independent	$C_5(u_1, u_2) = u_1 * u_2$	N/A

Note: Here Φ_2 denotes the standard bivariate normal CDF with correlation coefficient γ and Φ^{-1} is the inverse of the standard normal CDF

C is given by $F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$, where F_1 and F_2 are CDFs of the univariate binary distributions of Y_1 and Y_2 respectively. Following Panagiotelis et al. (2012), we can recover the joint probability mass function (PMF) of Y from the CDF as

$$P(Y_1 = y_1, Y_2 = y_2) = C(F_1(y_1), F_2(y_2); \theta) - C(F_1(y_1 - 1), F_2(y_2); \theta)$$
$$- C(F_1(y_1), F_2(y_2 - 1); \theta) + C(F_1(y_1 - 1), F_2(y_2 - 1); \theta). \tag{1}$$

The $C(u_1, u_2; \theta)$ could be any copula C_i, $1 \leq i \leq 5$, given in Table 1. The copula parameter θ is the correlation coefficient γ for the Gaussian copula, and it is α for the Clayton, Frank, and Gumbel copulas.

Selecting $y_i = 0, 1$ and noting that $F_i(0) = P(Y_i = 0) = 1 - p_i = q_i$ for $i = 1, 2$, Eq. (1) simplifies to the probabilities given in Table 2.

Given $\rho = \text{Corr}(Y_1, Y_2)$, and if we use in Table 2 the Gaussian copula C_1, the parameter $\theta = \gamma$ can be obtained by solving equation

$$\text{Corr}(Y_1, Y_2) = \rho = \frac{C_1(p_1, p_2; \gamma) - p_1 p_2}{\sqrt{p_1 q_1 p_2 q_2}}, \tag{2}$$

since $1 - q_1 - q_2 + C_1(q_1, q_2; \gamma) = C_1(p_1, p_2; \gamma)$.

Trivariate binary distributions

In this section we extend the pair-copula method to construct three dimensional binary distributions. Let $Y = (Y_1, Y_2, Y_3)$ be a vector of three correlated binary random variables. Note that

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = P(Y_2 = y_2) * P(Y_1 = y_1, Y_3 = y_3 | Y_2 = y_2). \tag{3}$$

The above equation shows the three dimension distribution can be obtained by constructing the bivariate conditional distribution of (Y_1, Y_3) given $Y_2 = y_2$. To this end we introduce some notation. We first construct bivariate distributions for (Y_1, Y_2) and (Y_2, Y_3) selecting bivariate copulas $C_{12}(u_1, u_2)$ and $C_{23}(u_1, u_2)$ from Table 2. For notational convenience we omit the copula parameter here and in some formulas later.

Table 2 PMF of bivariate binary variables

(Y_1, Y_2)	Probability
(0, 0)	$C(q_1, q_2; \theta)$
(0, 1)	$q_1 - C(q_1, q_2; \theta)$
(1, 0)	$q_2 - C(q_1, q_2; \theta)$
(1, 1)	$1 - q_1 - q_2 + C(q_1, q_2; \theta)$

Note: There are 3 parameters for this distribution, marginal means p_1, p_2 and copula parameter θ

Table 3 Conditional PMF of (Y_1, Y_3) given Y_2

	Probability
$(Y_1, Y_3)\|Y_2 = 0$	
(0, 0)	$C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0})$
(0, 1)	$q_{1\|0} - C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0})$
(1, 0)	$q_{3\|0} - C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0})$
(1, 1)	$1 - q_{1\|0} - q_{3\|0} + C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0})$
$(Y_1, Y_3)\|Y_2 = 1$	
(0, 0)	$C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1})$
(0, 1)	$q_{1\|1} - C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1})$
(1, 0)	$q_{3\|1} - C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1})$
(1, 1)	$1 - q_{1\|1} - q_{3\|1} + C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1})$

Note: $P(Y_1, Y_3|Y_2 = 0)$ has 6 parameters; marginal means p_1, p_2, p_3 and copula parameters $\theta_{12}, \theta_{23}, \theta_{13|Y_2=0}$. $P(Y_1, Y_3|Y_2 = 1)$ needs the same first five parameters and $\theta_{13|Y_2=1}$

Let $q_{1|0} = P(Y_1 = 0|Y_2 = 0) = \frac{C_{12}(q_1, q_2)}{q_2}$ and $p_{1|0} = 1 - q_{1|0} = 1 - \frac{C_{12}(q_1, q_2)}{q_2}$. Thus $Y_1|Y_2 = 0$ is distributed as Bernoulli with mean $p_{1|0}$. Similarly, $Y_3|Y_2 = 0$ is distributed as Bernoulli with mean $p_{3|0}$, where $p_{3|0} = 1 - \frac{C_{23}(q_2, q_3)}{q_2}$. We also have $Y_1|Y_2 = 1$ is Bernoulli with mean $p_{1|1} = 1 - \frac{q_1 - C_{12}(q_1, q_2)}{p_2}$, and $Y_3|Y_2 = 1$ is Bernoulli with mean $p_{3|1} = 1 - \frac{q_3 - C_{23}(q_2, q_3)}{p_2}$.

Table 3 shows the conditional distributions of $(Y_1, Y_3)|Y_2 = 0$ and $(Y_1, Y_3)|Y_2 = 1$. Finally, from Eq. (3) and using the conditonal distributions we can get the joint trivariate PMF as given in Table 4.

We give six numerical examples to see that different copulas and parameter values give rise to different trivariate binary distributions. All these six distributions have the same marginal means $p_1 = 0.8, p_2 = 0.7$ and $p_3 = 0.6$. The choice of the copulas and parameter values are summarized in Table 5 for these six cases.

We give details of the calculations only for case 5, the others are similarly done. In this case we start with Table 2 using Gaussian copula with $\gamma_{12} = 0.752, \gamma_{23} = 0.607$, and $p = (0.8, 0.7, 0.6)$. The resulting PMFs of bivariate binary variables are in Table 6.

Now, to construct the conditional PMFs of bivariate variables we need to get the marginal parameters of the conditional Bernoulli variables $Y_1|Y_2 = 0$ and $Y_1|Y_2 = 1$. These are

Table 4 PMF of trivariate binary variables

(Y_1, Y_2, Y_3)	Probability
(0, 0, 0)	$q_2 * C_{13\|0}(q_{1\|0}, q_{3\|0})$
(0, 0, 1)	$C_{12}(q_1, q_2) - q_2 * C_{13\|0}(q_{1\|0}, q_{3\|0})$
(0, 1, 0)	$p_2 * C_{13\|1}(q_{1\|1}, q_{3\|1})$
(0, 1, 1)	$q_1 - C_{12}(q_1, q_2) - p_2 * C_{13\|1}(q_{1\|1}, q_{3\|1})$
(1, 0, 0)	$C_{23}(q_2, q_3) - q_2 * C_{13\|0}(q_{1\|0}, q_{3\|0})$
(1, 0, 1)	$q_2 - C_{23}(q_2, q_3) - C_{12}(q_1, q_2) + q_2 * C_{13\|0}(q_{1\|0}, q_{3\|0})$
(1, 1, 0)	$q_3 - C_{23}(q_2, q_3) - p_2 * C_{13\|1}(q_{1\|1}, q_{3\|1})$
(1, 1, 1)	$1 - q_1 - q_2 - q_3 + C_{12}(q_1, q_2) + C_{23}(q_2, q_3) + p_2 * C_{13\|1}(q_{1\|1}, q_{3\|1})$

Note: There are 7 parameters here: marginal means p_1, p_2, p_3 and copula parameters $\theta_{12}, \theta_{23}, \theta_{13|Y_2=0}$, and $\theta_{13|Y_2=1}$

Table 5 Summary of parameter values for PMF of the trivariate binary variables

Case	Pair-copulas	Dependence
1	All Gaussian	$\gamma_{12} = 0.752, \gamma_{23} = 0.607$
		$\gamma_{13\|Y_2=0} = 0.480, \gamma_{13\|Y_2=1} = 0.233$
2	All Clayton	$\alpha_{12} = 2, \alpha_{23} = 1.5$
		$\alpha_{13\|Y_2=0} = \alpha_{13\|Y_2=1} = 0.4$
3	All Frank	$\alpha_{12} = \alpha_{23} = 1.85$
		$\alpha_{13\|Y_2=0} = 0.95, \alpha_{13\|Y_2=1} = 0.85$
4	All Gumbel	$\alpha_{12} = \alpha_{23} = 10$
		$\alpha_{13\|Y_2=0} = \alpha_{13\|Y_2=1} = 4$
5	Gaussian for marginals	$\gamma_{12} = 0.752, \gamma_{23} = 0.607$
	Frank for conditionals	$\alpha_{13\|Y_2=0} = 0.95, \alpha_{13\|Y_2=1} = 0.85$
6	All Independent	

$$q_{1|0} = \frac{P(Y_1 = 0, Y_2 = 0)}{q_2} = \frac{0.1517}{0.3} = 0.5057,$$

$$q_{3|0} = \frac{P(Y_2 = 0, Y_3 = 0)}{q_2} = \frac{0.2097}{0.3} = 0.6990,$$

$$q_{1|1} = \frac{P(Y_1 = 0, Y_2 = 1)}{p_2} = \frac{0.0483}{0.7} = 0.0690,$$

$$q_{3|1} = \frac{P(Y_2 = 1, Y_3 = 0)}{p_2} = \frac{0.1903}{0.7} = 0.2719.$$

Then, $p_{1|0} = 1 - 0.5057 = 0.4943$, $p_{3|0} = 1 - 0.6990 = 0.3010$, $p_{1|1} = 1 - 0.0690 = 0.931$ and $p_{3|1} = 1 - 0.2719 = 0.7281$. Also, Frank copula is used for the conditional distributions with parameters $\alpha_{13|Y_2=0} = 0.95$, $\alpha_{13|Y_2=1} = 0.85$. The PMFs of conditional bivariate binary variables are calculated according to Table 3, resulting in the values given in Table 7.

Since $P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = P(Y_2 = y_2) * P(Y_1 = y_1, Y_3 = y_3 | Y_2 = y_2)$, the last step is to multiply values in Table 7 by $P(Y_2 = y_2)$ to get the joint trivariate probability. For example, $P(Y_1 = 0, Y_2 = 0, Y_3 = 0) = 0.3782 * 0.3 = 0.1135$ or $P(Y_1 = 1, \ Y_2 = 1, \ Y_3 = 0) = 0.2475 * 0.7 = 0.1732$. The three dimensional joint binary distributions for the six cases are given in Table 8.

Table 6 PMF of bivariate binary variables using Gaussian copula

	Probability
(Y_1, Y_2)	
(0, 0)	$C(q_1, q_2; \gamma_{12}) = \Phi_2(\Phi^{-1}(0.2), \Phi^{-1}(0.3); 0.752) = 0.1517$
(0, 1)	$q_1 - C(q_1, q_2; \gamma_{12}) = 0.2 - 0.1517 = 0.0483$
(1, 0)	$q_2 - C(q_1, q_2; \gamma_{12}) = 0.3 - 0.1517 = 0.1483$
(1, 1)	$1 - q_1 - q_2 + C(q_1, q_2; \gamma_{12}) = 1 - 0.2 - 0.3 + 0.1517 = 0.6517$
(Y_2, Y_3)	
(0, 0)	$C(q_2, q_3; \gamma_{23}) = \Phi_2(\Phi^{-1}(0.3), \Phi^{-1}(0.4); 0.607) = 0.2097$
(0, 1)	$q_2 - C(q_2, q_3; \gamma_{23}) = 0.3 - 0.2097 = 0.0903$
(1, 0)	$q_3 - C(q_2, q_3; \gamma_{23}) = 0.4 - 0.2097 = 0.1903$
(1, 1)	$1 - q_2 - q_3 + C(q_2, q_3; \gamma_{23}) = 1 - 0.3 - 0.4 + 0.2097 = 0.5097$

Table 7 Conditional PMF of bivariate binary variables generated by Frank copula

	Probability
$(Y_1, Y_3 \| Y_2 = 0)$	
$(0, 0)$	$C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0})$ $= -\frac{1}{0.95} \log\left(1 + \frac{(e^{-0.95*0.5057}-1)(e^{-0.95*0.699}-1)}{e^{-0.95}-1}\right) = 0.3782$
$(0, 1)$	$q_{1\|0} - C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0}) = 0.5057 - 0.3782 = 0.1275$
$(1, 0)$	$q_{3\|0} - C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0}) = 0.699 - 0.3782 = 0.3208$
$(1, 1)$	$1 - q_{1\|0} - q_{3\|0} + C_{13\|0}(q_{1\|0}, q_{3\|0}; \theta_{13\|Y_2=0}) = 0.1735$
$(Y_1, Y_3 \| Y_2 = 1)$	
$(0, 0)$	$C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1})$ $= -\frac{1}{0.85} \log\left(1 + \frac{(e^{-0.85*0.069}-1)(e^{-0.85*0.2719}-1)}{e^{-0.85}-1}\right) = 0.0244$
$(0, 1)$	$q_{1\|1} - C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1}) = 0.069 - 0.0244 = 0.0446$
$(1, 0)$	$q_{3\|1} - C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1}) = 0.2719 - 0.0244 = 0.2475$
$(1, 1)$	$1 - q_{1\|1} - q_{3\|1} + C_{13\|1}(q_{1\|1}, q_{3\|1}; \theta_{13\|Y_2=1}) = 0.6835$

Comparison of pair-copula and MP models

In this section, we will compare the probability mass functions generated by the pair-copula methods and by the multivariate probit model. We will see that even if we use bivariate Gaussian copulas, these two methods yield different probability mass functions. Furthermore, the pair-copula method is successful in cases where the MP model fails to generate a probability mass function with specified univariate marginals and correlations.

Multivariate probit (MP) model

Let $Y = (Y_1, Y_2, \ldots, Y_m)$ be a vector of binary random variables. The multivariate probit model assumes that associated with the vector Y there is a latent vector $Z = (Z_1, Z_2, \ldots, Z_m)$, which is distributed as multivariate normal (MVN), such that $Y_t = 1$ if $Z_t > 0$, and $Y_t = 0$ if $Z_t \leq 0$. Assume $Z_t = \mu_t + \epsilon_t$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$ is MVN($\mathbf{0}, R$). Then, $p_t = P(Y_t = 1) = P(Z_t > 0) = P(\mu_t + \epsilon_t > 0) = \Phi(\mu_t)$, and $q_t = (1 - p_t) = \Phi(-\mu_t)$. The joint PMF of $Y = (Y_1, Y_2, \ldots, Y_m)$ is given by

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \int_{D_m} \cdots \int_{D_1} \frac{1}{(2\pi)^{\frac{m}{2}} |R|^{\frac{1}{2}}} \exp\left(-\frac{\boldsymbol{\epsilon}\, R^{-1}\, \boldsymbol{\epsilon}'}{2}\right) \, d\boldsymbol{\epsilon},$$

(4)

where $D_t = (-\infty, \mu_t)$ if $y_t = 1$, and $D_t = (\mu_t, \infty)$ if $y_t = 0$. For example, for $m = 3$ we have

Table 8 PMF of trivariate binary variables

(Y_1, Y_2, Y_3)	PMF_1	PMF_2	PMF_3	PMF_4	PMF_5	PMF_6
$(0, 0, 0)$	0.1267	0.1366	0.0582	0.1983	0.1135	0.024
$(0, 0, 1)$	0.0250	0.0322	0.0337	0.0000	0.0382	0.036
$(0, 1, 0)$	0.0210	0.0190	0.0450	0.0017	0.0171	0.056
$(0, 1, 1)$	0.0273	0.0122	0.0631	0.0000	0.0312	0.084
$(1, 0, 0)$	0.0830	0.0939	0.1077	0.0994	0.0963	0.096
$(1, 0, 1)$	0.0653	0.0373	0.1004	0.0023	0.0520	0.144
$(1, 1, 0)$	0.1693	0.1505	0.1891	0.1006	0.1732	0.224
$(1, 1, 1)$	0.4824	0.5182	0.4028	0.5977	0.4785	0.336

$$P(Y_1 = 0, Y_2 = 0, Y_3 = 0) = \int_{\mu_1}^{\infty} \int_{\mu_2}^{\infty} \int_{\mu_3}^{\infty} \phi_3(\epsilon_1, \epsilon_2, \epsilon_3 ; R) \ d\boldsymbol{\epsilon}$$

$$= \Phi_3(-\mu_1, -\mu_2, -\mu_3 ; R),$$

where $\Phi_3(\boldsymbol{\epsilon} ; R)$ is the CDF of trivariate standard normal with correlation matrix R.

Distributions generated by pair-copula and MP models

Since the MP model relies on the Gaussian distribution, for a fair comparison we will use the Gaussian copulas for the bivariate and conditional distributions in the pair-copula construction. Consider the case of two dimensions. In this case, taking C_{12} as the bivariate Gaussian copula, the PMF as given in Table 2 is $P(Y_1 = 0, Y_2 = 0) = C_{12}(q_1, q_2; \gamma) = \Phi_2(\Phi^{-1}(q_1), \Phi^{-1}(q_2); \gamma) = \Phi_2(-\mu_1, -\mu_2; \gamma)$, which is identical to the probability under the MP model. Therefore, the probability distributions are the same for two dimensions. For three dimensions for the MP model, we have

$$P(Y_1 = 0, Y_2 = 0, Y_3 = 0) = \Phi_3(-\mu_1, -\mu_2, -\mu_3 ; R). \tag{5}$$

From Table 4, we see that for the pair-copula model

$$P(Y_1 = 0, \ Y_2 = 0, \ Y_3 = 0) = q_2 \, C_{13|0}(q_{1|0}, q_{3|0}). \tag{6}$$

With bivariate Gaussian copulas we have $q_i = \Phi(-\mu_i)$ for $i = 1, 2, 3$ and

$$q_{1|0} = \frac{C_{12}(q_1, q_2)}{q_2} = \frac{\Phi_2(-\mu_1, -\mu_2; \gamma_{12})}{\Phi(-\mu_2)} = P(\epsilon_1 < -\mu_1 | \epsilon_2 < -\mu_2),$$

$$q_{3|0} = \frac{C_{23}(q_2, q_3)}{q_2} = \frac{\Phi_2(-\mu_2, -\mu_3; \gamma_{23})}{\Phi(-\mu_2)} = P(\epsilon_3 < -\mu_3 | \epsilon_2 < -\mu_2), \tag{7}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)$ is distributed as a standard trivariate normal with correlation matrix $R = (\gamma_{ij})$. The quantities $q_{1|0}$ and $q_{3|0}$ are the same as the corresponding values for the probit model. Taking $C_{13|0}(u_1, u_2)$ as bivariate Gaussian copula with correlation $\gamma_{13|0}$, Eq. (6) is equivalent to

$$P(Y_1 = 0, \ Y_2 = 0, \ Y_3 = 0) = \Phi(-\mu_2) \, \Phi_2\Big(\Phi^{-1}(q_{1|0}), \Phi^{-1}(q_{3|0}); \gamma_{13|0}\Big). \tag{8}$$

Clearly the quantity (8) is not equal to (5) since the parameter $\gamma_{13|0}$ can be any value in $(-1, 1)$ and need not be related to R. Thus the PMF of the pair-coupla model is different from the MP model.

An advantage of the pair-copula method over the MP model

In this section, we give an example to show that the pair-copula method is useful to construct multivariate binary distributions with specified marginals and correlation structure in cases where the multivariate probit model breaks down. Let's assume the marginal means are given by the vector $p = (0.2, \ 0.3, \ 0.2)$. For the equicorrelated structure the feasible range of the correlation parameter ρ is $(-0.25, 0.7638)$, see Theorem 1 in Chaganty and Joe (2006). For the value $\rho = 0.76$, the latent correlation matrix obtained by solving Eq. (2), for all pairs is

$$R = \begin{bmatrix} 1 & 0.9853 & 0.9411 \\ 0.9853 & 1 & 0.9853 \\ 0.9411 & 0.9853 & 1 \end{bmatrix},$$

Table 9 Input values for construction of the PMF of equicorrelated binary variables

	Input values
(Y_1, Y_2)	$p_1 = 0.2, p_2 = 0.3, \gamma_{12} = 0.9853$
(Y_2, Y_3)	$p_2 = 0.3, p_3 = 0.2, \gamma_{23} = 0.9853$
$(Y_1, Y_3\|Y_2)$	$p_{1\|0} = 0.001, p_{3\|0} = 0.001, \gamma_{13\|Y_2=0} = 0.8968$
	$p_{1\|1} = 0.6643, p_{3\|1} = 0.6643, \gamma_{13\|Y_2=1} = 0.6398$

which is not positive definite and thus the MP method does not give a PMF for the binary variables. However, the pair-copula method generates a PMF for the binary variables with specified marginal means and equicorrelated structure. The input values needed to calculate the PMF are listed in Table 9.

Proceeding as in "Trivariate binary distributions" section, the resulting three dimensional distribution is given in Table 10. We can check that this distribution has the specified marginal means $p_1 = 0.2$, $p_2 = 0.3$, and $p_3 = 0.2$, and equicorrelated structure with $\rho = 0.76$.

Extensions to four and higher dimensions

The pair-copula method for three dimensions described in "Trivariate binary distributions" section can be extended to construct four or higher-dimensional multivariate binary distributions. The foundations of these higher dimensional extensions have been laid out in Joe (1996; 1997). In a pioneering work, Bedford and Cooke (2002) showed how to use graphical models consisting of vines with trees and edges. The edges in a given tree become the nodes of the next tree. The vine structures not only help in enumerating and organizing numerous decompositions of a multivariate distribution but also facilitate models for different types of dependence for the marginal and conditional distributions. In recent years several articles have been published in the literature on vine pair-copula models, see Kurowicka and Joe (2011). The papers by Min and Czado (2010), Gruber and Czado (2015), and Dalla Valle et al. (2018) discuss Bayesian inference for these vine pair-copula models. The lecture notes by Czado (2019) and Brechmann and Schepsmeier (2013) discusses the practical implementation of vine copulas using the R software. The two most popular vines are the canonical C-vine and the drawable D-vine, see Czado (2019) and Joe (2014). An application of the C-vine for analyzing familial data is in Deng and Chaganty (2021). In this paper, we focus on the D-vine which is a natural candidate for analyzing longitudinal data which consists of an ordered sequence of variables.

Figure 1 shows the nested tree structure of the D-vine for m variables. There are $(m-1)$ trees and for the ith tree there are $m - i + 1$ nodes, represented by rectangular boxes. In the case of $m = 4$, the D-vine consists of 3 trees. For the pair-copula construction we will need bivariate copulas for the pairs (12), (23), and (34) in tree 1. Since we are dealing with binary variables, for tree 2 we will need two bivariate copulas for constructing the

Table 10 Three dimensional distribution with specified marginals

(Y_1, Y_2, Y_3)	Probability	(Y_1, Y_2, Y_3)	Probability
(0, 0, 0)	0.6989	(1, 0, 0)	0.0004
(0, 0, 1)	0.0004	(1, 0, 1)	0.0003
(0, 1, 0)	0.0627	(1, 1, 0)	0.0380
(0, 1, 1)	0.0380	(1, 1, 1)	0.1613
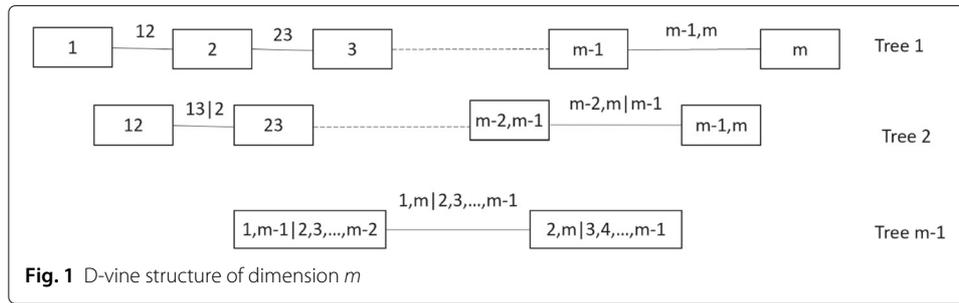
Fig. 1 D-vine structure of dimension *m*

conditional distribution of (13|2) and another two for constructing (24|3). The final tree requires the construction of the conditional distribution of (14|23), which in turn requires four bivariate copulas for the four possible values of the conditioned variables 2 and 3. The joint PMF in four dimensions is given by

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) = P(Y_1 = y_1, Y_4 = y_4 | Y_2 = y_2, Y_3 = y_3)$$
$$\times P(Y_2 = y_2, Y_3 = y_3), \tag{9}$$

The probability $P(Y_1 = y_1, Y_4 = y_4 | Y_2 = y_2, Y_3 = y_3)$ requires $p_{1|00} = P(Y_1 = 1 | Y_2 = 0, Y_3 = 0)$, $p_{1|01} = P(Y_1 = 1 | Y_2 = 0, Y_3 = 1)$,..., $p_{4|11} = P(Y_4 = 1 | Y_2 = 1, Y_3 = 1)$, which can be obtained from the bivariate and trivariate distributions constructed as in "Construction of vine pair-copula binary distributions" section.

Binary distributions with structured correlation matrices

To allow parsimonious modeling, multivariate binary distributions with structured correlation matrices are normally employed in the analysis of longitudinal or clustered binary data. The two most popular structured correlation matrices are autoregressive of order one (AR(1)) and equicorrelated. Yang and Chaganty (2014) have outlined a method of constructing a multivariate binary distribution with AR(1) structure, and here we focus on the equicorrelated structure.

In "An advantage of the pair-copula method over the MP model" section, we gave an example of a three-dimensional binary distribution with specified marginals and equicorrelated structure. The pair-copula method with bivariate Gaussian copulas can be used to generate higher-dimensional multivariate binary distributions with specified marginals and equicorrelated structure. This requires specification of the partial correlations $\text{Corr}(Y_1, Y_3 | Y_2)$, $\text{Corr}(Y_2, Y_4 | Y_3)$, ..., $\text{Corr}(Y_{m-2}, Y_m | Y_{m-1})$ for tree 2; $\text{Corr}(Y_1, Y_4 | Y_2, Y_3)$, ..., $\text{Corr}(Y_{m-3}, Y_m | Y_{m-2}, Y_{m-1})$ for tree 3; ....; $\text{Corr}(Y_1, Y_m | Y_2, ...Y_{m-1})$ for tree $m-1$. For binary variables, these partial correlations depend on the values of the conditional variables. To simplify matters we set $\text{Corr}(Y_i, Y_{i+k} | Y_{i+1}, ...Y_{i+k-1}) = \rho/(1 + (k-1)\rho)$. The motivation for this assumption comes from the result that for equicorrelated structure, partial correlation $\rho_{i,i+k|i+1,...,i+k-1}$ equals $\rho/(1 + (k-1)\rho)$ as shown in the Appendix. The corresponding parameter $\gamma_{i,i+k|i+1,...,i+k-1}$ of the bivariate Gaussian copula can be obtained by solving Eq. (2) using the two conditional probabilities $p_{i|i+1,...,i+k-1}$ and $p_{i+k|i+1,...,i+k-1}$. In the next section we give a numerical example to illustrate this method for dimension $m = 4$.

Table 11 Input values for construction of the PMF of equicorrelated binary variables

	Input values
(Y_1, Y_2)	$p_1 = 0.26, p_2 = 0.36, \gamma_{12} = 0.6156$
(Y_2, Y_3)	$p_2 = 0.36, p_3 = 0.25, \gamma_{23} = 0.6191$
(Y_3, Y_4)	$p_3 = 0.25, p_4 = 0.24, \gamma_{34} = 0.6214$
$(Y_1, Y_3 \mid Y_2)$	$p_{1\mid 0} = 0.1284, p_{3\mid 0} = 0.1201, \gamma_{13 \mid Y_2=0} = 0.5316$
	$p_{1\mid 1} = 0.4939, p_{3\mid 1} = 0.4809, \gamma_{13 \mid Y_2=1} = 0.4340$
$(Y_2, Y_4 \mid Y_3)$	$p_{2\mid 0} = 0.2491, p_{4\mid 0} = 0.1414, \gamma_{24 \mid Y_3=0} = 0.5049$
	$p_{2\mid 1} = 0.6926, p_{4\mid 1} = 0.5359, \gamma_{24 \mid Y_3=1} = 0.4531$
$(Y_1, Y_4 \mid Y_2, Y_3)$	$p_{1\mid 00} = 0.0931, p_{4\mid 00} = 0.0840, \gamma_{14 \mid Y_1=0, Y_3=0} = 0.4723$
	$p_{1\mid 01} = 0.3871, p_{4\mid 01} = 0.3220, \gamma_{14 \mid Y_1=0, Y_3=1} = 0.3538$
	$p_{1\mid 10} = 0.3564, p_{4\mid 10} = 0.3142, \gamma_{14 \mid Y_1=1, Y_3=0} = 0.3560$
	$p_{1\mid 11} = 0.6423, p_{4\mid 11} = 0.6308, \gamma_{14 \mid Y_1=1, Y_3=1} = 0.3510$

Numerical example of equicorrelated binary distribution

Assuming the marginal means are $p = (0.26, \ 0.36, \ 0.25, \ 0.24)$, the feasible range of the correlation parameter ρ is $(-0.3244, 0.7492)$ for the equicorrelated structure. Let $\rho = 0.4$, the distribution is calculated and presented in Table 12 using input values from Table 11.

We can check the marginal means of the distribution in Table 12 are $P(Y_1 = 1) = 0.26$, $P(Y_2 = 1) = 0.36$, $P(Y_3 = 1) = 0.25$, $P(Y_4 = 1) = 0.24$, and further the distribution has an equicorrelated structure with $\rho = 0.4$.

Parameter estimation

In this section, we discuss estimation of the parameters via maximum likelihood estimation (MLE) for the D-vine pair-copula model with bivariate Gaussian distributions. Suppose that there are n independent subjects, and there are m repeated binary observations on each subject. Thus the data consists of binary vectors $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{im})$ of dimension m. Let p_j be the marginal probability of y_{ij} assumed to be the same for all i. There are 2^m possible combinations for $\mathbf{y}_i$. For instance, when $m = 4$, we have 16 combinations, that is, $\mathbf{y}_i = (0,0,0,0)$, or $(0,0,0,1)$, or $(0,0,1,0)$, $\cdots$, or $(1,1,1,1)$. The n observations can be grouped into 2^m counts. Assume the number of $(0, \cdots, 0)$ vectors is n_1, the number of $(0, \cdots, 1)$ is n_2, so on and so forth, the number of $(1, \cdots, 1)$ is n_{2^m}. Using these notations, the loglikelihood, $\ell(\boldsymbol{\theta})$, for D-vine pair-copula model for a sample of n independent observations is given by

Table 12 Four dimensional distribution with specified marginals and equicorrelated structure

(Y_1, Y_2, Y_3, Y_4)	Probability	(Y_1, Y_2, Y_3, Y_4)	Probability
(0, 0, 0, 0)	0.4779	(1, 0, 0, 0)	0.0379
(0, 0, 0, 1)	0.0328	(1, 0, 0, 1)	0.0145
(0, 0, 1, 0)	0.0358	(1, 0, 1, 0)	0.0163
(0, 0, 1, 1)	0.0113	(1, 0, 1, 1)	0.0135
(0, 1, 0, 0)	0.0917	(1, 1, 0, 0)	0.0364
(0, 1, 0, 1)	0.0285	(1, 1, 0, 1)	0.0302
(0, 1, 1, 0)	0.0318	(1, 1, 1, 0)	0.0322
(0, 1, 1, 1)	0.0302	(1, 1, 1, 1)	0.0791

$$\ell(\boldsymbol{\theta}) \propto n_1 \log P(Y_{i1} = 0, Y_{i2} = 0, \dots, Y_{im} = 0) + n_2 \log P(Y_{i1} = 0, Y_{i2} = 0, \dots, Y_{im} = 1)$$
$$+ \dots + n_{2^m} \log P(Y_{i1} = 1, Y_{i2} = 1, \dots, Y_{im} = 1), \tag{10}$$

where the parameter $\boldsymbol{\theta}$ consists of marginal probabilities and copula parameters that are functions of correlations between the binary variables. Take the two dimensional example shown in Table 2 for instance, the loglikelihood is

$$\ell(\gamma_{12}, p_1, p_2) \propto n_1 \log(P(Y_{i1} = 0, Y_{i2} = 0)) + n_2 \log(P(Y_{i1} = 0, Y_{i2} = 1))$$
$$+ n_3 \log(P(Y_{i1} = 1, Y_{i2} = 0)) + n_4 \log(P(Y_{i1} = 1, Y_{i2} = 1))$$
$$= n_1 \log(C_1(q_1, q_2; \gamma_{12})) + n_2 \log(q_1 - C_1(q_1, q_2; \gamma_{12}))$$
$$+ n_3 \log(q_2 - C_1(q_1, q_2; \gamma_{12})) + n_4 \log(1 - q_1 - q_2 + C_1(q_1, q_2; \gamma_{12})), \tag{11}$$

where C_1 is the bivariate Gaussian copula. The maximum likelihood estimates of the parameters are obtained by maximizing (10) using the optimization routine "L-BFGS-B" by Byrd et al. (1995) which allows box constraints. The standard errors of the parameters are obtained from the Hessian matrix at optimized values using "Richardson" method of the function "Hessian" in the R package "numDeriv" by Gilbert and Varadhan (2012).

Data analysis
Here we present a real-life data analysis to illustrate the application of the D-vine pair-copula with bivariate Gaussian distributions. We also compare the results with the MP model and the model that ignores the correlation between the variables.

Drug response data
This data was first reported by Grizzle et al. (1969). Here 46 subjects were treated with three drugs 1, 2 and 3, and recorded their response as 0 for unfavorable or 1 for favorable. For example, (0, 0, 0) stands for unfavorable responses for all the three drugs. We assume the three binary responses are equicorrelated with correlation parameter ρ. The maximum likelihood estimates (MLE) of the marginal probabilities p_1, p_2 and p_3 and ρ together with standard errors (SE) are presented in Table 13.

The estimate of ρ is close to zero both for the MP and D-vine pair-copula models. The estimates and standard errors of D-vine independent copula model are listed at the last two columns of Table 13. The D-Vine independent copula model has the minimum AIC and seems to be a good choice for this data.

Table 13 Parameter estimates and standard errors for the drug response data

Parameter	MP			D-vine pair-copula			Independent D-Vine		
	MLE	SE	*p*-value	MLE	SE	*p*-value	MLE	SE	*p*-value
p_1	0.3911	0.0871	< 0.0001	0.3906	0.0725	< 0.0001	0.3913	0.0719	< 0.0001
p_2	0.4081	0.0994	< 0.0001	0.3905	0.0725	< 0.0001	0.3913	0.0719	< 0.0001
p_3	0.6635	0.1042	< 0.0001	0.6495	0.0700	< 0.0001	0.6522	0.0702	< 0.0001
ρ	0.0683	0.0216	0.0008	0.0395	0.0848	0.3207	NA	NA	NA
AIC	190.59			190.37			188.59		

Discussion

In recent years vine pair-copula models have become popular for analyzing dependent multivariate data. However, understanding and using these models for discrete in particular for binary data can pose as a challenge to the practitioner. In this paper, we have illustrated the pair-copula construction of binary distributions in the case of two and three dimensions that make it easy for the practitioner. In three dimensions using bivariate Gaussian copula, we have shown that the probability mass function generated by the pair-copula differs from the mass function of the multivariate probit (MP) model. We gave a numerical example where the MP model fails but one is able to use the pair-copula method to generate mass function with specified marginals and correlations. For four and higher dimensions we provide a method of constructing a multivariate binary distribution with specified marginals and equicorrelated structure using the D-vine pair-copula method. We discussed the maximum likelihood estimation of the parameters for grouped multivariate binary data and provided a real-life data analysis. Future work involves including covariates in these models.

Appendix

Consider the equicorrelated structure given by $R = (1 - \rho)I_m + \rho\, e_m e_m^T$, with parameter ρ. Here I_m is the identity matrix of dimension m and e_m is a $m \times 1$ column vector of ones. From formula (2.19), page 40 in Joe (2014), we have the partial correlation is given by

$$\rho_{1,m|2,\ldots,m-1} = \frac{\rho - \rho^2\, e_{m-2}^T R_{11}^{-1} e_{m-2}}{1 - \rho^2\, e_{m-2}^T R_{11}^{-1} e_{m-2}}, \tag{12}$$

where $R_{11} = (1 - \rho)I_{m-2} + \rho\, e_{m-2} e_{m-2}^T$. Using the formula in Example 4.1 of Chaganty (1997), we have

$$R_{11}^{-1} = \frac{1}{1-\rho}\left[I_{m-2} - \frac{\rho}{1 + (m-3)\rho} e_{m-2} e_{m-2}^T \right].$$

Since $e_{m-2}^T e_{m-2} = (m-2)$ we get

$$\begin{aligned}
e_{m-2}^T R_{11}^{-1} e_{m-2} &= \frac{1}{1-\rho}\left((m-2) - \frac{\rho}{1 + (m-3)\rho}(m-2)^2 \right) \\
&= \frac{(m-2)}{1-\rho}\left(1 - \frac{\rho(m-2)}{1 + (m-3)\rho} \right) \\
&= \frac{(m-2)}{1 + (m-3)\rho}.
\end{aligned} \tag{13}$$

Substituting (13) in (12) and simplfying we get

$$\begin{aligned}
\rho_{1,m|2,\ldots,m-1} &= \frac{\rho(1 + (m-3)\rho) - (m-2)\rho^2}{1 + (m-3)\rho - (m-2)\rho^2} \\
&= \frac{\rho - \rho^2}{m\rho(1-\rho) + (1-\rho)(1-2\rho)} \\
&= \frac{\rho}{1 + (m-2)\rho}.
\end{aligned} \tag{14}$$

The constant $(m-2)$ in the denominator of (14) represents the number of conditional variables. More generally, for the equicorrelated structure the partial correlation $\rho_{i,i+k|i+1,\ldots,i+k-1} = \rho/(1 + (k-1)\rho)$ for any $1 \leq i \leq (m-k)$, $1 \leq k \leq (m-1)$.

Declarations

References

Bedford, T., Cooke, R. M.: Vines–a new graphical model for dependent random variables. Ann. Stat. **30**(4), 1031–1068 (2002)

Brechmann, E., Schepsmeier, U.: Modeling dependence with c- and d-vine copulas: the r package cdvine. J. Stat. Softw., 52 (2013). https://doi.org/10.18637/jss.v052.i03

Byrd, R. H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)

Chaganty, N. R.: An alternative approach to the analysis of longitudinal data via generalized estimating equations. J. Stat. Plan. Inf. **63**(1), 39–54 (1997)

Chaganty, N. R., Joe, H.: Range of correlation matrices for dependent bernoulli random variables. Biometrika. **93**(1), 197–206 (2006)

Czado, C.: Analyzing Dependent Data with Vine Copulas: A Practical Guide With R. Springer International Publishing, Lecture Notes in Statistics (2019)

Dalla Valle, L., Leisen, F., Rossini, L.: Bayesian non-parametric conditional copula estimation of twin data. J. Roy. Stat. Soc. C: Appl. Stat. **67**, 523–548 (2018)

Deng, Y., Chaganty, N. R.: Pair-copula models for analyzing family data. J. Stat. Theory Pract. **15**(1), 13 (2021)

Escarela, G., Perez-Ruiz, L. C., Bowater, R. J.: A copula-based markov chain model for the analysis of binary longitudinal data. J. Appl. Stat. **36**(6), 647–657 (2009)

Gilbert, P., Varadhan, R.: numDeriv: Accurate Numerical Derivatives. R Package (2012). http://CRAN.R-project.org/package=numDeriv

Grizzle, J. E., Starmer, C. F., Koch, G. G.: Analysis of categorical data by linear models. Biometrics, 489–504 (1969)

Gruber, L., Czado, C.: Sequential bayesian model selection of regular vine copulas. Bayesian Anal. **10**, 937–963 (2015)

Joe, H.: Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. Lecture Notes–Monograph Series, vol. 28. Institute of Mathematical Statistics, Hayward (1996)

Joe, H.: Multivariate Models and Multivariate Dependence Concepts. Chapman & Hall/CRC, London (1997)

Joe, H.: Dependence modeling with copulas. Chapman and Hall/CRC, London (2014)

Kurowicka, D., Joe, H.: Dependence modeling: vine copula handbook. World scientific, Singapore (2011)

Lennon, H.: Gaussian copula modelling for integer-valued time series. PhD thesis. The University of Manchester (United Kingdom) (2016)

Liang, K. Y., Zeger, S. L.: Longitudinal data analysis using generalized linear models. Biometrika. **73**(1), 13–22 (1986)

Min, A., Czado, C.: Bayesian inference for multivariate copulas using pair-copula constructions. J. Financ. Econ. **8**, 511–546 (2010)

Panagiotelis, A., Czado, C., Joe, H.: Pair copula constructions for multivariate discrete data. J. Am. Stat. Assoc. **107**(499), 1063–1072 (2012)

Panagiotelis, A., Czado, C., Joe, H., Stöber, J.: Model selection for discrete regular vine copulas. Comput. Stat. Data Anal. **106**, 138–152 (2017)

Radice, R., Marra, G., Wojtyś, M.: Copula regression spline models for binary outcomes. Stat. Comput. **26**(5), 981–995 (2016)

Sabo, R. T., Chaganty, N. R.: What can go wrong when ignoring correlation bounds in the use of generalized estimating equations. Stat. Med. **29**(24), 2501–2507 (2010)

Sklar, M.: Fonctions de repartition an dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris. **8**, 229–231 (1959)

Smith, M., Min, A., Almeida, C., Czado, C.: Modeling longitudinal data using a pair-copula decomposition of serial dependence. J. Am. Stat. Assoc. **105**(492), 1467–1479 (2010)

Smith, M. S., Khaled, M. A.: Estimation of copula models with discrete margins via bayesian data augmentation. J. Am. Stat. Assoc. **107**(497), 290–303 (2012)

Winkelmann, R.: Copula bivariate probit models: with an application to medical expenditures. Health Econ. **21**(12), 1444–1455 (2012)

Yang, W., Chaganty, N. R.: A contrasting study of likelihood methods for the analysis of longitudinal binary data. Commun. Stat. Theory Methods. **43**(14), 3027–3046 (2014)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.