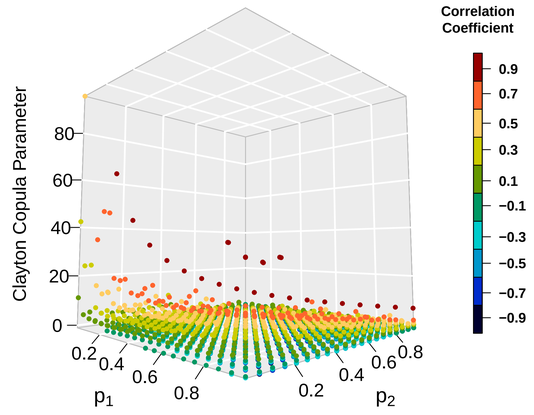
D-Vine Copula Model For Dependent Binary Data

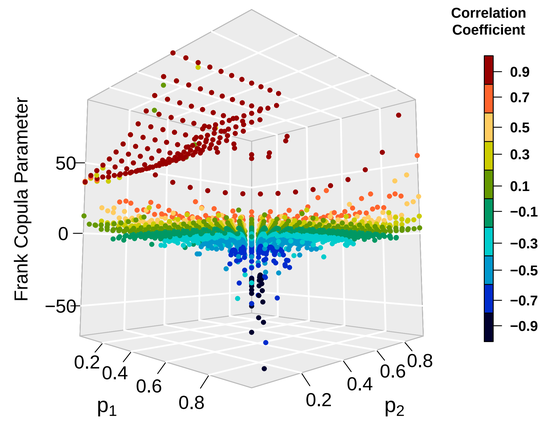
Huihui Lin (hlin005@odu.edu) and N. Rao Chaganty (rchagant@odu.edu)
Department of Mathematics and Statistics
Old Dominion University

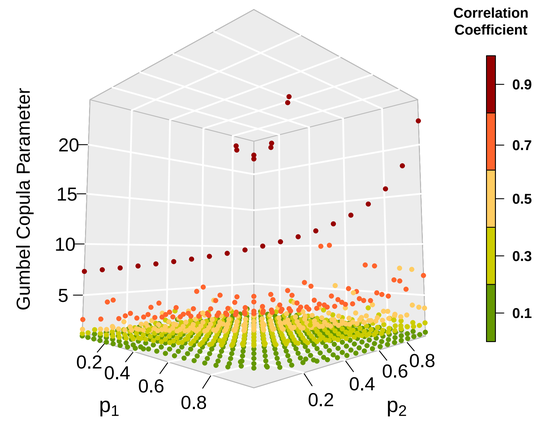
Introduction

High-dimensional dependent binary data are prevalent in a wide range of scientific disciplines. A popular method for analyzing such data is the Multivariate Probit (MP) model [1]. But the MP model sometimes fails even within a feasible range of binary correlations, because the underlying correlation matrix of the latent variables may not be positive definite. In this research we proposed pair copula models, assuming the dependence between the binary variables is first order autoregressive (AR(1))or equicorrelated structure.
The outline of this poster presentation is as follows. We start with the definition of the copula and pictorially illustrate the relation between the copula parameter and the binary correlaton. We illustrate pair copula constructions of multivariate binary distributions using D-vines and C-vines. We show the application of our method on a real life data. Finally, we briefly discuss our ongoing research.

Relationship between copula parameter and marginal proportions with binary correlation







D-vines and C-vines

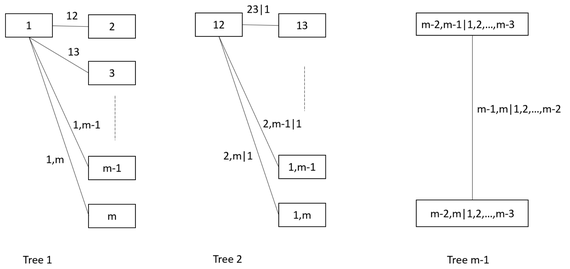


Figure 1: M-dimensional C-vine structure.

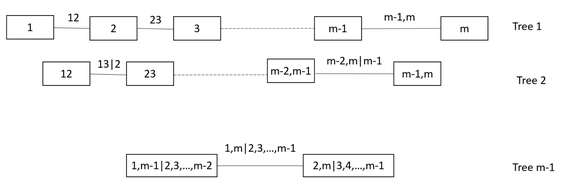


Figure 2: M-dimensional D-vine structure.

Copula Definition

A multivariate copula is a cumulative distribution function (CDF) with uniform (0,1) univariate marginals [2]. The bivariate copula is sufficient for our pair copula models because a multivariate copula can be constructed using a D-vine and bivariate copula margins. We denote a bivariate copula by $C_\theta(u_1, u_2)$, where θ is the correlation coefficient if the copula is Gaussian, otherwise it is simply a parameter for Clayton, Gumbel and Frank copulas.

Bivariate Binary Variables

(y_1, y_2)	Probability
(0, 0)	$C_\theta(q_1, q_2)$
(0, 1)	$q_1 - C_\theta(q_1, q_2)$
(1, 0)	$q_2 - C_\theta(q_1, q_2)$
(1, 1)	$1 - q_1 - q_2 + C_\theta(q_1, q_2)$

Trivariate Binary Variables

Let $q_{1|0} = P(Y_1 = 0 \mid Y_2 = 0) = \frac{C_{12}(q_1,q_2)}{q_2}$, and $q_{1|1}$ similarly.

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

Ex.AR(1) Trivariate Binary Var.

(y_1, y_2, y_3)	MP	Gaussian	Clayton	Frank	Gumbel
(0, 0, 0)	0.1854	0.1846	0.1846	0.1846	0.1846
(0, 0, 1)	0.3516	0.3524	0.3524	0.3524	0.3524
(0, 1, 0)	0.0174	0.0182	0.0182	0.0182	0.0182
(0, 1, 1)	0.1154	0.1147	0.1147	0.1147	0.1147
(1, 0, 0)	0.0689	0.0697	0.0697	0.0697	0.0697
(1, 0, 1)	0.1339	0.1331	0.1331	0.1331	0.1331
(1, 1, 0)	0.0181	0.0173	0.0173	0.0173	0.0173
(1, 1, 1)	0.1088	0.1096	0.1096	0.1096	0.1096

NOTE: AR(1) coefficient is $\rho = 0.2$, marginal means are $p = (0.33,\ 0.26,\ 0.71)$.

Real Data Example

A randomized experiment was conducted by [3] to evaluate the impact of the driver education on the number of collisions and violations among teenage drivers. In our study, we analyze the dependence relationship using our pair copula models with only the follow-up data of the control group for four years.

	MP		Gaussian D-Vine	
Parameter	EST	SE	EST	SE
p_1	0.1673	0.0018	0.1539	0.0073
p_2	0.3288	0.0029	0.3030	0.0094
p_3	0.3777	0.0054	0.3743	0.0098
p_4	0.3824	0.0032	0.3959	0.0100
ρ	0.1429	0.0010	0.1672	0.0121
AIC	11277.85		11265.05	

The table above shows the maximum likelihood estimates, standard errors and AIC for the models. All the marginal probabilities and AR(1) correlation parameter ρ are very close in both models. From the analysis we can conclude that the traffic violations rate increases with time.

References

[1] Weiming Yang and N Rao Chaganty. A contrasting study of likelihood methods for the analysis of longitudinal binary data. *Communications in Statistics-Theory and Methods*, 43(14):3027–3046, 2014.

[2] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[3] John R Stock, JK Weaver, HW Ray, JR Brink, Michael G Sadof, et al. Evaluation of safe performance secondary school driver education curriculum demonstration project. Technical report, United States. National Highway Traffic Safety Administration, 1983.

Ongoing Research

Our work in progress involves developing D-vine pair Gaussian copula regression models for binary longitudinal data assuming AR(1) structure.