Computational Statistics and Data Analysis 103 (2016) 28–55

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Copula in a multivariate mixed discrete–continuous model

Aurelius A. Zilko*, Dorota Kurowicka

*Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands*

ARTICLE INFO

*Article history:*
Received 11 June 2015
Received in revised form 15 February 2016
Accepted 15 February 2016
Available online 4 May 2016

*Keywords:*
Copula
Vine
Mixed models

ABSTRACT

The use of different copula-based models to represent the joint distribution of an eight-dimensional mixed discrete and continuous problem consisting of five discrete and three continuous variables is investigated. The discussion starts with the theoretical properties of the copula-based models. Four different models are constructed for the data collected for the purpose of predicting the length of disruption caused by problems with the train detection system in the Dutch railway network and their performance is tested. The more complex models turn out to represent the data better. Nevertheless, it is shown that the simpler eight dimensional Normal copula still constitutes a statistically sound model for the data.

## 0. Introduction

Copulas separate information present in the margins from the dependence in the joint distribution. They have been proven to be very attractive in many different applications where a joint distribution of continuous variables is of interest. However, when copulas are used for discrete models, Genest and Nešlehová (2007) show that the popular way of copula parameters estimation, through finding an empirical dependence measure and equating it to the theoretical one, is highly biased. Nevertheless, the maximum likelihood technique can still be used, even if it is much more computationally expensive.

Maximum likelihood estimation of copula parameters for discrete models requires an approximation of a multidimensional integral or evaluating $2^n$ finite differences of the copula to find the value of the probability mass function of an $n$-dimensional model. Due to computational costs, many copula applications of discrete models have only involved lower dimensional problems. Nikoloulopoulos and Karlis (2008) constructed a four-dimensional Bernoulli distribution with the help of several different copula families with three parameters and Song et al. (2009) built a trivariate discrete distribution with the Normal copula. In both cases, the copula models worked well and the authors highlighted that the dependence structure between the variables did not only come from the copula but also from the margins.

Nikoloulopoulos (2013) proposed computing the rectangle probabilities using the simulated maximum likelihood approach method. The new approach has been shown in Nikoloulopoulos (2015) to be applicable in dimension of up to 225, even though as dimension and sample size increase, computational burden becomes heavy. Another alternative technique to estimate the parameters uses the Bayesian methods as proposed by Smith and Khaled (2012). However, this technique is also computationally intensive.

The reduction in estimation cost of copula parameters has been achieved in Panagiotelis et al. (2012) by using the copula-vine approach. The multivariate discrete distribution has been constructed with a set of pairwise bivariate (conditional) copulas arranged according to a graphical structure called a regular vine (for more information about vines, see Kurowicka and Joe, 2011). The conditional copulas in this construction are assumed not to depend on the conditioning variables. The computation cost of calculating the probability mass function with this approach only grows as $2n(n-1)$, which makes this model applicable even for very high dimensional problems.

* Corresponding author.
*E-mail address:* A.A.Zilko@tudelft.nl (A.A. Zilko).

http://dx.doi.org/10.1016/j.csda.2016.02.017
0167-9473/© 2016 Elsevier B.V. All rights reserved.

Similarly to purely discrete models, mixed discrete and continuous models with copulas encounter problems. Most applications of copulas to low dimensional problems are available in the literature. Song et al. (2009) model a bivariate mixed binary discrete (disposition) and continuous (severity of burn injury) variables with a Normal copula. De Leon and Wu (2010) proposed two strategies to compute the maximum likelihood for a bivariate mixed discrete and continuous distribution with a simulation study and an application to the same data set as in Song et al. (2009). He et al. (2012) used the Normal copula to construct two and three dimensional mixed discrete and continuous models each with one discrete variable to study the relationship between the genotype (discrete) and a few continuous phenotypes such as the cholesterol density and the protein concentration. Stöber et al. (2015) constructed a six-dimensional mixed discrete and continuous model with five binary variables and one continuous variable representing six chronic diseases by following the copula-vine approach with constant conditional copulas as described in Panagiotelis et al. (2012).

In the first part of this paper, we concentrate on theoretical issues concerning the use of copulas for purely discrete and mixed discrete–continuous models. A few simple results of the existence of a copula model for the joint distribution of binary variables are provided. This investigation provides a background for the exploration of copula models for a mixed discrete and continuous data presented in Zilko et al. (2015), where five binary and three continuous variables are used to construct a latency time model that is part of the railway disruption length model. The goal is to choose a model that allows fast and accurate prediction of the latency time for different combinations of values of the other variables in the model.

The rest of the paper is organized as follows. Section 1 introduces copula models for multivariate Bernoulli distribution. In Section 2 mixed discrete–continuous models with copulas are presented. Section 3 is concerned with the application of copula models to the latency time data. This section contains the results. Finally, conclusions and short discussions on how the model that is constructed in this paper will be used in practice are presented in Section 4.

## 1. Multivariate Bernoulli distribution with copulas

The aim of this section is to lay a theoretical background and discuss copula models for discrete and mixed discrete–continuous distributions. We start with the multivariate Bernoulli distribution and investigate the existence of a copula family that allows representation of such a distribution. We present a copula construction that allows to model any multivariate Bernoulli distribution.

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector taking values in $\{0, 1\}^n$ and $\mathbf{x} = (x_1, \dots, x_n)$ be a realization of $\mathbf{X}$. The joint probability is

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= p(x_1, x_2, \dots, x_n)\\ &= p(0, 0, \dots, 0)^{\prod_{j=1}^{n}(1-x_j)} p(1, 0, \dots, 0)^{x_1 \prod_{j=2}^{n}(1-x_j)} \dots p(1, 1, \dots, 1)^{\prod_{j=1}^{n} x_j}\end{aligned} \tag{1}$$

where all the $p$'s must sum up to 1. The marginal distribution of $X_i$ is

$$\mathbb{P}(X_i = 0) = p_i = \sum_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n \in \{0,1\}} p(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n).$$

Another popular representation of a multivariate Bernoulli distribution is the log-linear expansion. Taking the logarithm of the probability in (1) and collecting the appropriate terms leads to:

$$\begin{aligned}\log p(x_1, x_2, \dots, x_n) = {} & \log p(0, 0, \dots, 0) + \sum_i u_i x_i + \sum_{i,j} u_{ij} x_i x_j\\ & + \sum_{ijk} u_{ijk} x_i x_j x_k + \dots + u_{12\dots n} x_1 x_2 \dots x_n.\end{aligned} \tag{2}$$

The $u$-terms in (2) represent the two, three, ..., $n$-way interactions between the variables (see e.g. Whittaker, 1990) and they can be obtained from the probabilities as follows:

$$\begin{aligned} u_1 &= \log \frac{p(1, 0, 0, \dots, 0)}{p(0, 0, 0, \dots, 0)},\\ u_{12} &= \log \frac{p(1, 1, 0, \dots, 0)p(0, 0, 0, \dots, 0)}{p(1, 0, 0, \dots, 0)p(0, 1, 0, \dots, 0)},\\ u_{123} &= \log \frac{p(1, 1, 1, 0, \dots, 0)p(1, 0, 0, 0, \dots, 0)p(0, 1, 0, 0, \dots, 0)p(0, 0, 1, 0, \dots, 0)}{p(1, 1, 0, 0, \dots, 0)p(1, 0, 1, 0, \dots, 0)p(0, 1, 1, 0, \dots, 0)p(0, 0, 0, 0, \dots, 0)}.\end{aligned} \tag{3}$$

The interactions between the variables contain information about dependence. The term $u_{12}$ is also known as the log cross-product ratio ($cpr$) between variables $X_1$ and $X_2$. Notice that the cross product ratio $cpr(X_1, X_2)$ can be rewritten in terms of conditional probabilities of variables $X_1$ and $X_2$ given all remaining variables $X_3, \dots, X_n$ equal zero. Moreover, $u_{123}$