Journal of Multivariate Analysis 110 (2012) 106–120

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

# In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes

V. Radu Craiu *, Avideh Sabeti

*Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario M5S 3G3, Canada*

ARTICLE INFO

*Article history:*
Available online 2 April 2012

*AMS subject classifications:*
62H05
62F15
62C40

*Keywords:*
Adaptive Markov chain Monte Carlo
Bayesian inference
Conditional copula model
Cubic spline
Deviance information criterion
Mixed outcomes

ABSTRACT

Conditional copula models are flexible tools for modelling complex dependence structures in regression settings. We construct Bayesian inference for the conditional copula model adapted to regression settings in which the bivariate outcome is continuous or mixed. The dependence between the copula parameter and the covariate is modelled using cubic splines. The proposed joint Bayesian inference is carried out using adaptive Markov chain Monte Carlo sampling. The deviance information criterion (DIC) is used for selecting the copula family that best approximates the data and for choosing the calibration function. The performances of the estimation and model selection methods are investigated using simulations.



## 1. Introduction

Central to modern statistical analysis is modelling and understanding the dependence between random variables. However, the number of options at the statistician's disposal is limited due to the sparsity of available multivariate distributions. Copulas represent a flexible alternative in which one can bypass the use of multivariate distributions by using Sklar's theorem [48] to model separately the marginal distributions and the joint dependence structure.

In this paper we consider the bivariate case, as the extension to more than two responses requires a considerably more complex statistical machinery. Let $Y_1$ and $Y_2$ be continuous random variables of interest with joint distribution function $H$ and marginal distributions $F_1$ and $F_2$, respectively. Sklar's theorem ensures the existence of a unique copula $C : [0, 1]^2 \rightarrow [0, 1]$, which satisfies $H(y_1, y_2) = C\{F_1(y_1), F_2(y_2)\}$, for all $(y_1, y_2) \in \mathbb{R}^2$. Often, the function $C$ assumes a parametric form indexed by a copula parameter $\theta$. A nice introduction to copulas and their properties is [38] and the connections between various copulas and dependence concepts are discussed in detail in [31].

It should be noted that copula models have been mostly used for dependence between continuous random variables (e.g., [9,10,14,17,28,46,59]) but there is growing interest in the study and applications of copula models for mixed (discrete and continuous) data [21,13,42,52,53,55]. Recent work by [21,49] has shown that extra care is needed in performing and interpreting statistical inference for copula models when some of the marginals are discrete.

While frequentist methods have been used predominantly in the copula literature (see, for instance, [20,22]), the availability of powerful computers and "off-the-shelf" algorithms have recently led to a few Bayesian methods for copula estimation, model selection and goodness-of-fit (see, for instance, [29,34,37,42,47,49]).

A natural extension of the classical copula model allows the copula parameter to vary with covariate values as in [35]. This idea, formalized by the conditional copula model of [40], allows for realistic copula modelling in regression settings

* Corresponding author.
*E-mail addresses:* craiu@utstat.toronto.edu (V.R. Craiu), sabeti@utstat.toronto.edu (A. Sabeti).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.03.010

[3,23]. For instance, if, in addition to $Y_1$ and $Y_2$, we have information on a covariate $X$, then the influence of $X$ on the dependence between $Y_1$ and $Y_2$ can be modelled by the conditional copula $C(\cdot \mid X)$, which is the joint distribution function of $U_1 \equiv F_{1|X}(Y_1 \mid x)$ and $U_2 \equiv F_{2|X}(Y_2 \mid x)$ given $X = x$, where $Y_i \mid X = x$ has cdf $F_{i|X}(\cdot \mid x)$, $i = 1, 2$. For each $x$ in the support of $X$, it was showed in [40] that for continuous random variables the joint conditional distribution is uniquely defined by

$$H_X(y_1, y_2 \mid x) = C\{F_{1|X}(y_1 \mid x), F_{2|X}(y_2 \mid x) \mid x\}, \quad \text{for all } (y_1, y_2) \in \mathbb{R}^2.$$

In general, it is not known how the copula parameter varies with the covariate $X$. Moreover, visualizing the dependence patterns in the data is not possible as the available samples are not identically distributed (usually there is no replication at a given value of $X$). Therefore, one needs to use flexible methods to model the *calibration function* that characterizes the relationship between the copula parameter and the covariates. This naturally leads to the use of semiparametric [3] and nonparametric inferential tools [39,58] in the case of continuous marginals. Conditional copulas have been used mostly in the context of financial time series [5,32,40,41] but have recently been used in classical regression models too [3].

In this paper we propose joint Bayesian inference for a conditional copula model in which of interest, besides the marginal models, is the dependence between the outcomes. Our approach can handle the case of continuous and mixed (binary and continuous) outcomes. The Bayesian approach proposed here offers a number of advantages. First, it is a principled method to produce full likelihood-based inference as recommended by [21,49] in the case of copula models with some of the margins discrete. Second, the Bayesian approach offers a natural solution to the "propagation of errors" challenge in which the variance of the marginal model estimators must be accounted for when assessing the variance of the copula estimators. Within the Bayesian paradigm, the posterior distribution gives a full representation of the uncertainty in the *whole* data and prior. An added bonus is that the simultaneous estimation of both marginal distributions' parameters and copula parameters results in better understanding of the parameter dependences and leads to better performance of model selection criteria, as discussed in [47].

The shape of the calibration function is difficult to model in general and we thus need to use flexible models to capture its structure. In this paper we consider a Bayesian cubic spline model in which the choice for the position of the knots is data-driven. Sampling from the posterior distribution is performed using an adaptive MCMC algorithm following the principles developed in [26]. For more details regarding the theory and implementation of adaptive MCMC, we refer to [11,25,43].

In the next section we detail the statistical model along with the prior specification. Section 3 contains the computational algorithm used to sample from the posterior and the criterion used for model selection. The simulation study is summarized in Section 4. The paper closes with a discussion in which future directions are outlined.

## 2. The model

In this section we describe the conditional copula model for the dependent outcome data. We separate the model formulations used for continuous and for mixed outcomes (binary and continuous), as the two situations require different computational algorithms to sample from the posterior distribution and the resulting estimators exhibit different efficiency.

### 2.1. Continuous outcomes case

The data consist of bivariate continuous random variables $(V_1, V_2)$ and covariate $X$ measured for $n$ samples. Marginally, $V_i$ and $X$ are related through $V_i \sim \mathcal{N}(X\beta_i, \sigma_i^2)$, for $i = 1, 2$. The conditional dependence between $V_1$ and $V_2$ is defined via a conditional copula model with joint density

$$f(V_1, V_2|X) = \prod_{i=1}^{2} \frac{1}{\sigma_i}\phi\left(\frac{V_i - X\beta_i}{\sigma_i}\right) \times c^{(1,1)}\left\{\Phi\left(\frac{V_1 - X\beta_1}{\sigma_1}\right), \Phi\left(\frac{V_2 - X\beta_2}{\sigma_2}\right)\middle|\, \theta(X)\right\},$$

where $c^{(a,b)}(u, v|\theta) = \partial^{a+b}C(u, v|\theta)/\partial u^a \partial v^b$, for all $0 \le a, b \le 1$.

An important part of the model is the specification of $\theta(X)$. In this paper we assume that $X \in \mathbb{R}$ and follow [3,4] by assuming that $g(\theta) = \eta(X)$ where $g$ is a known function that maps the support of the copula parameter onto the real line and $\eta : \mathbb{R} \to \mathbb{R}$ is the unknown *calibration function* we want to estimate. We adopt the flexible cubic spline model suggested by [50] in which

$$\eta(z) = \sum_{j=0}^{3} \alpha_j z^j + \sum_{k=1}^{K} \psi_k (z - \gamma_k)_+^3, \tag{1}$$

where $a_+ = \max(0, a)$. It is well known that the performance of spline-based estimators are influenced by the location of the knots $\gamma_1, \ldots, \gamma_K$. In our model this choice is automatic and data-driven.

### 2.2. Mixed outcomes case

In the mixed outcome case, the response consist of one binary and one continuous random variable, denoted $Q$ and $W$, respectively. We are interested in statistical inference for the marginal logistic and linear regression models but also the estimation of the calibration function.