Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses

Q1 Jakob Stöber[a,*], Hyokyoung Grace Hong[b], Claudia Czado[a], Pulak Ghosh[c]

[a] *Center for Mathematical Sciences, Technische Universität München, Germany*

[b] *Department of Statistics and Probability, Michigan State University, United States*

[c] *Department of Quantitative Methods and Information systems, Indian Institute of Management, Bangalore, India*

ARTICLE INFO

Article history:
Received 26 April 2014
Received in revised form 1 February 2015
Accepted 2 February 2015
Available online xxxx

Keywords:
R-vine
Pair copula construction
GLM
LSOA II

ABSTRACT

Joint modeling of multiple health related random variables is essential to develop an understanding for the public health consequences of an aging population. This is particularly true for patients suffering from multiple chronic diseases. The contribution is to introduce a novel model for multivariate data where some response variables are discrete and some are continuous. It is based on pair copula constructions (PCCs) and has two major advantages over existing methodology. First, expressing the joint dependence structure in terms of bivariate copulas leads to a computationally advantageous expression for the likelihood function. This makes maximum likelihood estimation feasible for large multidimensional data sets. Second, different and possibly asymmetric bivariate (conditional) marginal distributions are allowed which is necessary to accurately describe the limiting behavior of conditional distributions for mixed discrete and continuous responses. The advantages and the favorable predictive performance of the model are demonstrated using data from the Second Longitudinal Study of Aging (LSOA II).

1. Introduction

The aim of this study is to demonstrate the use of a novel copula model for discrete and continuous response variables, which will help to broaden our understanding of pathways to comorbid conditions. We apply this model to data from the Second Longitudinal Study of Aging (LSOA II), which contains information on chronic diseases in the age group of 70+ on the national level.

The prevalence of chronic diseases tends to increase with age. Heart disease, stroke, hypertension, diabetes, obesity, and arthritis are among the most common. While the aforementioned conditions are often studied in an isolated setting, the elderly are likely to develop “comorbid conditions”, which refers to one or more diseases or conditions occurring together with the primary condition. Although there have been extensive studies exploring the relationship between two conditions controlling for other comorbid conditions, little research has been focused on comorbid conditions in a systematic joint modeling framework. This might be helpful to fill the gaps in our current understanding of comorbidity and reveal multivariate relationships.

Given the discrete nature of some response variables, copula models for continuous data cannot be applied to the LSOA II data. There are two standard methods for discrete marginal distributions in copula modeling. (i) For copula functions available in closed form, the probability mass function (pmf) can be computed by taking finite differences of the copula function

[*] Corresponding author.
E-mail address: stoeber@ma.tum.de (J. Stöber).

http://dx.doi.org/10.1016/j.csda.2015.02.001
0167-9473/© 2015 Published by Elsevier B.V.

for the discrete margins. This means that the number of evaluations of the copula function grows exponentially with the number of discrete variables (for our PCC model, the number of evaluations of copula functions only grows quadratically). Recent advances in computational capabilities and in approximation methods to the likelihood (see Masarotto and Varin, 2012 or Nikoloulopoulos, 2013) increase the scope of application for this method. However, the basic challenge that the computational complexity increases significantly with dimension and sample size remains. For further applications of models of this class see for example Shen and Weissfeld (2006), Nikoloulopoulos and Karlis (2006), Song et al. (2009) or He et al. (2012). (ii) As an alternative to the direct application of a copula to discrete data, latent continuous variables may be introduced. Then, the dependence structure of the latent variables is modeled instead of the discrete variables (see Pitt et al., 2006; Hoff, 2007; Dobra and Lenkoski, 2011; Murray et al., 2013, where this approach is applied for Gaussian models, Smith and Khaled, 2012, Danaher and Smith, 2011 extend the approach to a non-Gaussian setup). This has appealing features since it enables practitioners to apply well-known dependence models and also helps to avoid technicalities when working with discrete copulas (Nešlehová, 2007; Genest and Nešlehová, 2007). However, inference for such models is usually computationally difficult due to the latent variables.

The method presented here is based on pair copula constructions (PCCs) and has two major advantages over existing copula models. By generalizing the models of Panagiotelis et al. (2012) and Aas et al. (2009), it is computationally efficient for discrete variables and makes maximum likelihood inference feasible in high dimensions. It further combines different and also asymmetric copula families in a multivariate model, giving rise to very flexible higher dimensional distributions.

The remainder of the paper is structured as follows. Section 2 introduces the multivariate model which we consider, and inference and model selection is considered in Section 3. The motivating data set of our study is analyzed in Section 4. Section 5 summarizes our results and concludes the paper.

2. Multivariate model

In this section, we introduce the basic model using GLMs and the copula paradigm. In a generic form, let Y_{ijt} be the response/outcome of the ith patient for chronic disease j at observation/wave t, with $i = 1, 2, \ldots, N, j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T$. The covariates we consider in our analysis for patient i, disease j and time observation t are accordingly denoted as $\mathbf{x}_{ijt}$.

For all j, t, we assume that Y_{ijt} are independent and have distribution function

$$F_j(y_{ijt}|\mu_{ijt}, \phi_{j,t}),$$

where the mean parameter $\mu_{ijt} = h_j(\mathbf{x}_{ijt}\boldsymbol{\beta}_{jt}^T)$ is a function of the covariates and ϕ_{jt} is a possible scaling parameter. In particular, for j corresponding to a continuous response variable (the BMI in the data set which we will consider later), F_j can be the inverse Gaussian distribution with distribution function

$$F_{ig}(y|\mu, \phi) = \Phi\left(\sqrt{\frac{\phi}{y}}\left(\frac{y}{\mu} - 1\right)\right) + e^{\frac{2\phi}{\mu}}\,\Phi\left(-\sqrt{\frac{\lambda}{y}}\left(\frac{y}{\mu} + 1\right)\right),$$

and h_j can be chosen as $h_j(\cdot) = \exp(\cdot)$. If j corresponds to a binary response variable indicating the presence/absence of a chronic disease, a natural choice for F_j is the Bernoulli cdf with

$$F_b(y|\mu) = \begin{cases} 1 & y \geq 1 \\ 1 - \mu & 1 > y \geq 0 \\ 0 & 0 > y. \end{cases}$$

Here, the canonical choice for the link function h_j is $h_j = \frac{1}{1+e^{-(\cdot)}}$.

Furthermore, we assume that for any t, the marginal distributions F_j are linked with a copula function C_t. Hence, the joint distribution function for the outcome variables $(Y_{i,1,t}, \ldots, Y_{i,J,t})$ given covariates $(\mathbf{x}_{i1t}, \ldots, \mathbf{x}_{iJt})$ is given as

$$F_t(y_{i,1,t}, y_{i,2,t}, \ldots, y_{i,J,t}|\mathbf{x}_{i1t}, \ldots, \mathbf{x}_{iJt}) = C_t(F_1(y_{i,1,t}|\mu_{i1t}, \phi_{1t}), F_2(y_{i,2,t}|\mu_{i2t}, \phi_{2t}), \ldots, F_J(y_{i,J,t}|\mu_{iJt}, \phi_{Jt})). \quad (1)$$

This copula function is constructed from pair copula functions by subsequent conditioning. To illustrate the general principle, let us first consider a three dimensional example with two continuous variables $Y_1 \in \mathbb{R}$, $Y_3 \in \mathbb{R}$ with densities f_1, f_3 and one discrete variable $Y_2 \in \mathbb{Z}$ with pmf p_2. For the decomposition into bivariate building blocks, we start with the (generalized) joint density of $\mathbf{Y} = (Y_1, Y_2, Y_3)$. With *generalized* density, we mean the density of $\mathbf{Y}$ w.r.t. the product measure on the respective supports of the marginal variables. For discrete margins with values in $\mathbb{R}$ this is the counting measure on the set of possible outcomes, for continuous margins we consider the Lebesgue measure in $\mathbb{R}$. Given the cumulative distribution function $F_{\mathbf{Y}}$ of $\mathbf{Y}$, it is given by

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = \frac{\partial^2}{\partial y_1 \partial y_3}\left(F_{\mathbf{Y}}(y_1, y_2, y_3) - F_{\mathbf{Y}}(y_1, y_2 - 1, y_3)\right),$$

while the generalized density f_2 of Y_2 is its pmf $f_2(\cdot) = p_2(\cdot)$. By conditioning, the joint density can be decomposed as follows:

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = f_{1|2,3}(y_1|y_2, y_3) \cdot f_{2|3}(y_2|y_3) \cdot f_3(y_3). \quad (2)$$

Please cite this article in press as: Stöber, J., et al., Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.02.001