Comorbidity of chronic diseases in the elderly: Patterns Identified by a Copula Design for Mixed Responses

Jakob Stöber * Hyokyoung Grace Hong †

Claudia Czado* Pulak Ghosh ‡

*Center for Mathematical Sciences, Technische Universität München, Germany. Corresponding author email: stoeber@ma.tum.de

†Assistant Professor, Department of Statistics and Probability, Michigan State University

‡Professor, Department of Quantitative Methods and Information systems, Indian Institute of Management, Bangalore

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).

Keywords: R-vine, pair copula construction, GLM, 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 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), D. Hoff (2007), Dobra and Lenkoski (2011), Murray et al. (2013), where this is 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 i-th 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 $h_j(.) = exp(.)$. 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 $h_j = \frac{1}{1+e^{-(.)}}$.

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

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

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 $f_2(.) = p_2(.)$. By conditioning, the joint density can be decomposed as follows:

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

Using Sklar's theorem, we will now decompose the conditional densities in (2). Let us first consider the distribution of Y_1 and Y_3 given $Y_2 = y_2$ for some $y_2 \in \mathbb{Z}$, which has a

corresponding copula $C_{13|2}$. To simplify the following calculations and later inference, we will assume that $C_{13|2}$ does not depend on y_2. This means that we are working with a *simplified pair copula construction*, for a discussion in the continuous case see Stöber et al. (2013). For the conditional densities in (2), this means that

$$f_{1|2,3}(y_1|y_2,y_3) = c_{13|2}(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2)) \cdot f_{1|2}(y_1|y_2)$$
$$f_{2|3}(y_2|y_3) = \left(\frac{\partial}{\partial y_3} C_{23}(F_2(y_2), F_3(y_3)) - \frac{\partial}{\partial y_3} C_{23}(F_2(y_2-1), F_3(y_3))\right) \Big/ f_3(x_3),$$

where C_{23} is the copula corresponding to the bivariate marginal distribution of Y_2 and Y_3. Similarly, with the copula function C_{12} corresponding to the bivariate marginal distribution of Y_1 and Y_2, $f_{1|2}(y_1|y_2)$ can be further decomposed as

$$f_{1|2}(y_1|y_2) = \frac{\partial}{\partial y_1}\left(\frac{C_{12}(F_1(y_1), F_2(y_2)) - C_{12}(F_1(y_1), F_2(y_2-1))}{F_2(y_2) - F_2(y_2-1)}\right).$$

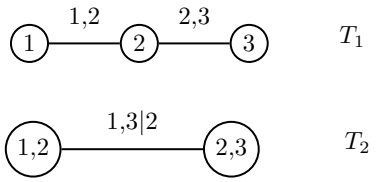
From this, $F_{1|2}(y_1|y_2)$ is easily obtained as

$$F_{1|2}(y_1|y_2) = \frac{C_{12}(F_1(y_1), F_2(y_2)) - C_{12}(F_1(y_1), F_2(y_2-1))}{F_2(y_2) - F_2(y_2-1)},$$

and the expression for $F_{3|2}(y_3|y_2)$ follows analogously. Thus, $f_{\mathbf{Y}}(y_1,y_2,y_3)$ can be expressed in terms of only the corresponding marginal distributions and the three bivariate copulas C_{12}, C_{23} and $C_{13|2}$. To illustrate this graphically, we can use two connected trees (Figure 1).

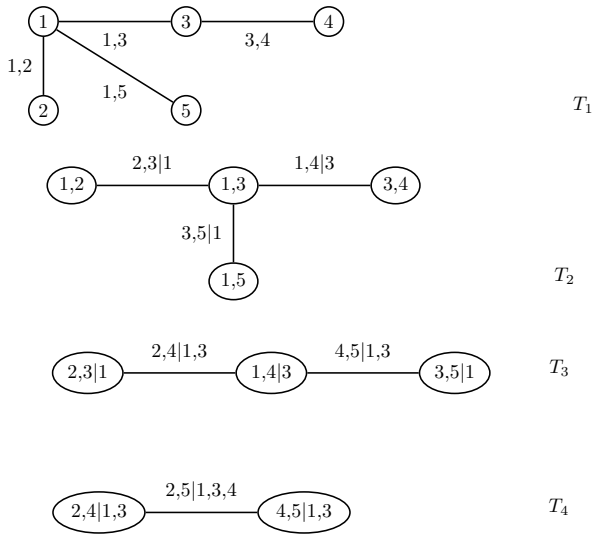
The first tree has the marginal variables as nodes and edges between 1 and 2 as well as between 2 and 3 to represent the copula functions C_{12} and C_{23}. The second tree contains the edges from the first tree as nodes and an edge between them to represent the conditional copula $C_{13|2}$.

Figure 1 The trees representing the three dimensional example. The edges correspond to copula functions in the decomposition.



The illustrated decomposition can be generalized to the d-dimensional case, with copulas corresponding to a d-dimensional analogue of the trees in Figure 1. For this, Bedford and Cooke (2001, 2002) introduced the regular vine (R-vine) as a graph theoretic tool to organize valid decompositions. In general, an R-vine on d variables constitutes a sequence of linked trees $\mathcal{V} = (T_1, \ldots, T_{d-1})$. As in our 3-dimensional example, the first tree has nodes $\{1, \ldots, d\}$ corresponding to the d variables, and $d-1$ edges corresponding to unconditional pair copulas. The nodes of each subsequent tree T_i are the edges of the previous tree T_{i-1}, and the edges correspond to pair copulas of bivariate distributions conditioned on $i-1$ variables in the decomposition. To make sure that the conditional cdfs which form the arguments of these conditional copulas can be calculated directly using already available copula functions (as for the arguments $F_{1|2}$ and $F_{3|2}$ of $C_{13|2}$ in our example) we require the *proximity condition* to hold: If two nodes in tree T_{i+1} are joined by an edge, the corresponding edges in T_i must share a common node. A five dimensional example is illustrated in Figure 2. Here, edges 3,4|2 and 2,5|3 share the common node 2,3 in tree T_2 and can thus be joined by an edge in tree T_3. Edges 3,4|2 and 2,5|3 both contain the numbers 2,3 of the common node. Thus, these numbers form the *conditioned* set of the new edge, while the remaining numbers 4,5 which are contained only for one of the nodes become the *conditioning* set. The new edge is labeled 4,5|2,3, it will correspond to the copula

Figure 2 An R-vine tree sequence in five dimensions with edge indices.



of the conditional distribution of Y_4 and Y_5 given $Y_2 = y_2, Y_3 = y_3$ in a distribution corresponding to the R-vine.

3 Inference and model selection

Now we outline the inference procedures which we will use to fit the described model to observed data. We will follow the paradigm of maximum-likelihood (ML) estimation to choose parameters to maximize the likelihood of observed data under the assumed model (for an overview of Bayesian methods for PCCs see Czado et al. (2013) and Smith et al. (2010)). Under standard regularity conditions, the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_n$ for n independent observations is strongly consistent and asymptotically normal:

$$\sqrt{n}\ I(\boldsymbol{\theta}_0)^{1/2}\Big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\Big) \xrightarrow{d} N(0, I_p),$$

where $\boldsymbol{\theta}_0$ is the true parameter and I_p the $p \times p$ identity matrix. The Fisher information matrix $I(\boldsymbol{\theta}_0)$ can be approximated by the observed information $I_n(\hat{\boldsymbol{\theta}})$ at the ML estimate $\hat{\boldsymbol{\theta}}_n$ defined as

$$I_n(\hat{\boldsymbol{\theta}}_n) = \left[\Big(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \sum_{i=1}^{n} \ell_n(\boldsymbol{\theta}) \Big)_{i,j=1,\dots,p} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}, \tag{3}$$

where $\ell_n(\boldsymbol{\theta})$ is the log likelihood function. This asymptotic behavior allows to approximate the covariance matrix of the parameter estimates $\hat{\boldsymbol{\theta}}_n$ by $\frac{1}{n} I_n(\hat{\boldsymbol{\theta}}_n)^{-1}$. Exploiting the hierarchical nature of R-vine copulas (see Stöber and Schepsmeier (2013)), the Hessian matrix in (3) can be calculated analytically for our model. This enables us to calculate the observed information and to estimate standard errors for copula parameter estimates as well as p-values for regression parameters.

To reduce the complexity, two-step estimation procedures are popular for copula models (Joe and Xu 1996; Joe 1997). Since our main interest is to estimate joint and conditional probabilities, we will follow this two-step approach in model selection and (i) choose covariates and interaction terms for the marginal models (Section 3.1) first and estimate the marginal parameters assuming independence and (ii) subsequently choose an appropriate PCC (Section 3.2). Once the model is selected, we re-estimate the marginal and copula parameters using joint ML estimation.

3.1 Marginal regression models

For the selection and initial parameter estimation for marginal regression models we use the statistical software package R (R Development Core Team 2011). To select the relevant covariates and interactions from a given set of possible covariates we will apply the Akaike Information Criterion (Akaike 1974), which is given by the negative log-likelihood plus the number of parameters as a punishment term.

To minimize this criterion, a stepwise procedure starting with a fully saturated model (i.e. including all possible covariates and interactions) is applied, removing in each step the term with the highest possible reduction in AIC until no further reduction is possible. To fit the parameters of the GLMs, we apply iteratively reweighted least squares for maximum likelihood estimation (Green 1984).

3.2 Regular vine copula

In this section, we start with calculating the generalized density of an R-vine copula model in arbitrary dimension. Based on this we illustrate how a suitable copula model is selected adopting ideas of Dißmann et al. (2013).

Let us consider d random variables $\mathbf{Y}_{1:d} = (Y_1, \dots, Y_d)$, with a joint density function $f(y_1, \dots, y_d)$. By subsequent conditioning, we can factorize this density as

$$f_{1:d}(y_1, \dots, y_d) = f_{1|2:d}(y_1|y_2, \dots, y_d) \cdot f_{2|3:d}(y_2|y_3, \dots, y_d) \cdot \ldots \cdot f_d(y_d). \tag{4}$$

This yields an expression where every term is of the form $f_{j|(j+1):d}(y_j|y_{j+1}, \dots, y_d)$, and we choose $j < h \leq d$ to further decompose (4). Let us denote $\mathbf{y}_{(j+1):d\setminus h} := (y_{j+1}, \dots, y_{h-1}, y_{h+1}, \dots, y_d)$. In the case where Y_j and Y_h are both continuous, we can express $f_{j|(j+1):d}$ as

$$\begin{aligned} f_{j|(j+1):d}(y_j|\mathbf{y}_{(j+1):d}) = f_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h}) \\ \cdot c_{j,h|(j+1):d\setminus h}\left(F_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})\right) \end{aligned} \tag{5}$$

by applying the bivariate version of Sklar's Theorem for densities. For discrete variables, we write $F_j(y_{j,0}) := F_j(y_j)$ and $F_j(y_{j,1})$ for the left-hand limit of F_j at y_j to simplify notation. In the case where $Y_j \in \mathbb{Z}$, this corresponds to $F_j(y_{j,1}) = F_j(y_j - 1)$.

With this, we can decompose the conditional density $f_{j|(j+1):d}(y_j|y_{j+1},\dots,y_d)$ as

$$\begin{aligned}
&f_{j|(j+1):d}(y_j|\mathbf{y}_{(j+1):d}) = P(Y_j = y_j|\mathbf{Y}_{(j+1):d} = \mathbf{y}_{(j+1):d}) \\
&= \frac{P(Y_j = y_j, Y_h = y_h|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})}{P(Y_h = y_h|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})} \\
&= \frac{\sum_{i_j,i_h=0,1}(-1)^{i_j+i_h}P(Y_j \le y_{j,i_j}, Y_h \le y_{h,i_h}|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})}{P(Y_h = y_h|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})} \\
&= \sum_{i_j,i_h=0}^{1}(-1)^{i_j+i_h}\frac{C_{j,h|(j+1):d\setminus h}\left(F_{j|(j+1):d\setminus h}(y_{j,i_j}|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_{h,i_h}|\mathbf{y}_{(j+1):d\setminus h})\right)}{f_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})} \\
&= c_{j,h|(j+1):d\setminus h}\cdot f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}),
\end{aligned} \tag{6}$$

c.f. (Panagiotelis et al. 2012), where we write

$$\begin{aligned}
&c_{j,h|(j+1):d\setminus h} := \\
&\sum_{i_j,i_h=0}^{1}(-1)^{i_j+i_h}\frac{C_{j,h|(j+1):d\setminus h}\left(F_{j|(j+1):d\setminus h}(y_{j,i_j}|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_{h,i_h}|\mathbf{y}_{(j+1):d\setminus h})\right)}{f_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h})},
\end{aligned} \tag{7}$$

for the discrete equivalent of the copula density in the continuous case. Let us now assume that Y_j is discrete and Y_h is continuous. We denote the derivative of a copula $C(\cdot,\cdot)$ with respect to its first (second) argument by $\partial_1 C(\cdot,\cdot)$ ($\partial_2 C(\cdot,\cdot)$). Defining

$$\begin{aligned}
&c_{j,h|(j+1):d\setminus h} := \\
&= \sum_{i_j=0}^{1}(-1)^{i_j}\frac{\partial_2 C_{j,h|(j+1):d\setminus h}(F_{j|(j+1):d\setminus h}(y_{j,i_j}|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h}))}{f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h})},
\end{aligned} \tag{8}$$

allows to write the conditional density $f_{j|(j+1):d}$ as

$$\begin{aligned}
&f_{j|(j+1):d}(y_j|\mathbf{y}_{(j+1):d}) = P(Y_j = y_j|\mathbf{Y}_{(j+1):d} = \mathbf{y}_{(j+1):d}) \\
&= \frac{\partial}{\partial y_h}F_{j,h|(j+1):d\setminus h}(y_{j,1}, y_h|\mathbf{y}_{(j+1):d\setminus h}) - \frac{\partial}{\partial y_h}F_{j,h|(j+1):d\setminus h}(y_{j,2}, y_h|\mathbf{y}_{(j+1):d\setminus h}) \\
&= \sum_{i_j=0}^{1}(-1)^{i_j}\partial_2 C_{j,h|(j+1):d\setminus h}(F_{j|(j+1):d\setminus h}(y_{j,i_j}|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})) \\
&= c_{j,h|(j+1):d\setminus h}\cdot f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}).
\end{aligned} \tag{9}$$

Similarly if Y_j is continuous and Y_h discrete, we write

$$\begin{aligned} &c_{j,h|(j+1):d\setminus h} := \\ &= \sum_{i_h=0}^{1} (-1)^{i_h} \frac{\partial_1 C_{j,h|(j+1):d\setminus h}(F_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_{h,i_h}|\mathbf{y}_{(j+1):d\setminus h}))}{f_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})}, \end{aligned} \tag{10}$$

and obtain

$$\begin{aligned} &f_{j|(j+1):d}(y_j|\mathbf{y}_{j(+1):d}) = \frac{\partial}{\partial y_j} F_{j|(j+1):d}(y_j|\mathbf{y}_{(j+1):d}) \\ &= \frac{\partial}{\partial y_j} \left[\frac{P(Y_j \leq y_j, Y_h = y_h|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})}{P(Y_h = y_h|\mathbf{Y}_{(j+1):d\setminus h} = \mathbf{y}_{(j+1):d\setminus h})} \right] \\ &= \sum_{i_h=0}^{1} (-1)^{i_h} \frac{\partial_1 C_{j,h|(j+1):d\setminus h}(F_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}), F_{h|(j+1):d\setminus h}(y_{h,i_h}|\mathbf{y}_{(j+1):d\setminus h}))}{f_{h|(j+1):d\setminus h}(y_h|\mathbf{y}_{(j+1):d\setminus h})} \\ &\quad \cdot f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}) = c_{j,h|(j+1):d\setminus h} \cdot f_{j|(j+1):d\setminus h}(y_j|\mathbf{y}_{(j+1):d\setminus h}). \end{aligned} \tag{11}$$

It is now clear that each term in (4) can be further decomposed into an expression containing appropriate bivariate copula functions, and conditional marginal densities. Further, it involves marginal conditional distribution functions as arguments for the bivariate copula terms. These can be obtained analogously to the marginal conditional densities above. Since the conditional marginal densities occurring in (5) - (11) can again be decomposed using the appropriate equation from (5) - (11), we conclude that by subsequent conditioning and application of Sklar's theorem, we can decompose the joint density $f_{1:d}(y_1, \ldots, y_d)$ of a multivariate random variable as a product of terms involving bivariate (pair-) copulas, acting on appropriate conditional distributions. However, given a decomposition into conditional densities (4), there are still many choices to make (we have to choose a corresponding h for each j), and also the ordering of the variables in (4) can be arbitrary.

As Bedford and Cooke (2001, 2002) showed, all these choices can be expressed by choosing a corresponding regular vine structure $\mathcal{V}$ with edge sets E_i, $i = 1, \ldots, d-1$.

3.2 Regular vine copula

The general expression for the (generalized) density of a general R-vine copula on which we base our inference procedures is

$$f_{1:d}(y_1, \ldots, y_d) = \prod_{i=1}^{d} f_i(y_i) \cdot \prod_{i=1}^{d-1} \prod_{e \in E_i} c_{j(e),k(e)|D_e}, \tag{12}$$

where we write $c_{j(e),k(e)|D_e}$ also for discrete variables as introduced before. As (12) shows, an R-vine copula with parametric components is specified by (i) an R-vine tree structure $\mathcal{V}$, (ii) the choice of a parametric pair copula family for each edge e and (iii) the corresponding parameter θ_e. While we will estimate the parameters in (iii) jointly with the parameters of the marginal regression models later, we must first select a suitable copula model, i.e. (i) choose a tree structure $\mathcal{V}$ and (ii) choose a suitable bivariate copula for each edge from a given set of available copulas.

As a first approach, we adopt the algorithm of Dißmann et al. (2013), which is modified for the case of mixed responses as follows: (i) For each pair of variables and each parametric pair copula family under consideration, calculate the corresponding value of the Akaike information criterion (AIC) from the copula data set. In our 3-dimensional example, we would calculate the AIC for the pairs $1, 3$, $2, 3$ and $1, 2$. (ii) Create a fully connected graph where the set of nodes N is the set of marginal variables (ex: $\{1, 2, 3\}$), and the set of edges E contains an edge between every possible pair of variables (ex: $1, 3$, $2, 3$ and $1, 2$). Associate to each edge the highest AIC value which has been estimated for the corresponding variables in step (i) as edge weight. (iii) Using the algorithm of Prim (1957) determine the maximum spanning tree corresponding to this graph, i.e. find a tree which maximizes the sum of edge weights (In our example this would have been the tree T_1 in Figure 1, i.e. the tree containing edges $1, 2$ and $2, 3$). (iv) For each edge in the resulting tree, choose the family for which we had obtained the highest AIC. (v) For each pair of edges $i, k|D$

and $i, j|D$ sharing a common node, determine pseudo observations for the next tree by applying the conditional distribution functions $F_{k|i,D}$ and $F_{j|i,D}$ to the data. Because of the proximity condition, these are all pseudo observations which might be required. (ex: $1, 2$ and $2, 3$ share 2, we compute $F_{1|2}$ and $F_{3|2}$) Proceed with the pseudo observations as in steps 1 to 4, while only considering edges which respect the proximity condition in step 2, until all trees together with their copula types and parameters are determined.

4 Application: Comorbidity in the elderly

4.1 Data description

Our motivating data comes from the Second Longitudinal Study of Aging (LSOA II), whose sample is nationally representative and is comprised of 9447 noninstitutionalized civilians in the US who were 70 years old and over at the time of the interview. Data in LSOA II were collected at three times: the baseline interview was done in 1994-1996 (Wave 1). The same subjects had two consecutive follow-up interviews between 1997 and 1998 (Wave 2), and between 1999 and 2000 (Wave 3). We note that the time gaps among consecutive interviews vary by individuals, but each interview was done about 2 years apart. LSOA II data, which is available from `http://www.cdc.gov/nchs/lsoa/lsoa2.htm`, provides valuable information on long term biomedical, social, and other various aspects of age-related conditions. Especially, chronic conditions were thoroughly asked during each follow-up, allowing us to explore the long term trajectory of these conditions. While serial dependence is likely to be present in this data, this is not explicitly described in our model where we focus on changes in the dependence structure we observe over the three wave. In future research, we can also consider explicitly describing the serial dependence of the outcomes. Among many chronic conditions, we focus on the occurrence of the

following six chronic conditions: hypertension (hyp), diabetes (dia), arthritis (art), heart disease (hd), stroke (str), and obesity/underweight via the body mass index (BMI). In general, although obesity causes more serious problems than underweight, underweight is also known to be a risk factor for many health conditions for the elderly. While often BMI is dichotomized by defining a binary outcome of obesity (yes or no) we analyze BMI as a continuous variable. This allows studying the association between the covariables and the entire distribution of the BMI, without losing the information imposed by its discretization (Fonseca et al. 2008). In particular, by this modeling approach we can appropriately study the relevance of underweight or possible benefits of moderate overweight which might help solve the ongoing controversy about appropriate BMI thresholds for elderly patients (Flicker et al. 2010; Singh et al. 2011). Information on the presence of the six chronic conditions was collected using standardized telephone interviews and self-administered questionnaires. In the baseline interview, the following questions were asked regarding the presence/absence of diseases: "Do you have XXX (a chronic disease)?" for hypertension and diabetes; "Ever had XXX?" for arthritis, heart disease (coronary heart disease, angina, heart attack, myocardial infarction), and stroke. During the follow-up studies, the subjects were inquired about their current conditions and asked "Had XXX since last interview?". Response categories were recoded as yes or no for each condition, except for BMI. Height was measured at the baseline, the BMI of each person at the different time points was then updated based on their current weight.

Out of 9447 subjects, 5294 had missing data due to death and unknown reasons in the follow-up interviews (or Wave 1 and Wave 2). Missing data can cause unbalanced data that cannot fit into the proposed model that requires all outcomes being fully observed. Therefore, if there are one or more components missing, we deleted the data. Although we treated the data as missing at random due to model complexity,

4.1 Data description

Table 1 Percentage of subjects with each condition who have another chronic condition. Diagonal contains subjects with that condition only. Percentages do not add to 100 because some patients have more than two conditions.

		Wave 2 No. (%) With Comorbid Condition					
Chronic condition	**No. of Subjects**	**Hyper-tension**	**Diabetes**	**Arthritis**	**Obesity**	**Heart Disease**	**Stroke**
Hypertension	1035	228(22)	147 (14.2)	682 (65.9)	157 (15.2)	220 (21.3)	55 (5.3)
Diabetes	237	147 (62)	22 (9.3)	156 (65.8)	57 (24.1)	71 (30)	17(7.2)
Arthritis	1470	682 (46.40)	156 (10.6)	523 (35.6)	195 (13.3)	323 (22)	69 (4.7)
Obesity	282	157(55.7)	57 (20.2)	195 (69.1)	31 (11)	45 (16)	8(2.80)
HD	441	220 (49.9)	71 (16.1)	323 (73.2)	45 (10.2)	50 (11.3)	40 (9.1)
Stroke	92	55 (59.8)	17 (18.5)	69 (75)	8 (8.7)	40 (43.5)	5(5.4)*

accounting dropout probability is encouraged for a more valid analysis (see Hong et al. (2013)). Finally, our dataset included a subsample of 2444 patients after removing missing information or "Don't know" responses. If someone responded "don't know" for one question, but responded "yes" or "no" to other conditions, they were excluded entirely. High rates of missing data among older people are not surprising and well known. Note that in our sample subjects who died during the survey period were not considered, since we focus on co-evolution of co-morbid conditions of the elderly. Therefore, our analysis focused on the subjects who survived until the end of the interview period between 1994 and 2000.

We find that comorbidity or multiple diseases were a common phenomenon in the elderly population. Table 1 (data for Wave 2 only, other waves available upon request) shows that arthritis and hypertension were the most common chronic conditions in the sample. However, among those who had arthritis, only about 35% had arthritis only. Similarly, about 20% of the subjects, who suffered from hypertension, had hypertension only. Most subjects with at least one chronic condition also suffered from arthritis or hypertension.

4.2 Predictor variables

Although there are many potentially useful predictors for our analysis, sex, age, income, education, and smoking are certainly among the most common and are, therefore, used as covariates for the marginal GLMs in our model (Table 2). The incidence of chronic diseases is known to increase with age; gender is associated with the progression and prevalence of chronic diseases. Further, Fleischer et al. (2011) reported association of a socioeconomic gradient for education and income with the risk factor profile for chronic diseases. People coping with chronic diseases are particularly vulnerable to the hazardous health effects of tobacco use. Smoking can exacerbate and complicate symptoms of the chronic conditions.

Table 2 Covariates included in the model

Covariate	Description
sex	$\{0 = female, 1 = male\}$, the sex of the subject
age	continuous, $70 - 95$, age at the beginning of the study
income	continuous, 0 - 26, income level of the subject
edu	continuous, 0 - 18 education level of the subject
smoke	$\{0 = non - smoker, 1 = smoker\}$

4.3 Joint model for the six response variables

The marginal distributions of the six response variables in our data set are described using GLMs. For the continuous distribution of the positive BMI values, we use an inverse Gaussian GLM with log-link, while the absence/presence of the chronic conditions is described by binomial GLMs with logit link. To decide which bivariate copula families to include in the R-vine copula selection procedure and to demonstrate the superior predictive performance of our joint copula model compared to independent regression models, we perform 10-fold cross-validation (see Arlot and Celisse (2010) for an overview of cross-validation procedures) as follows: The data is randomly par-

titioned into 10 patient sets of (almost) equal size. In each step, we leave out one of these subsets and apply the model selection procedures from Sections 3.1 and 3.2 for all three waves of observations. We include 10 different sets of pair copula families for the vine copula selection as shown in Table 3. More details on the bivariate copula families and parametrization we use can be found in Schepsmeier and Stöber (2014). The prediction quality of the resulting models for the remaining data set is then compared using the log predictive score (see Gneiting and Raftery (2007) for a review of scoring rules). Table 3 lists the sum of log predictive scores for the 10 subsets where we subtracted the scores corresponding to the benchmark independence model.

Table 3 Differences of log predictive scores to the independence models for different sets of copula families under consideration. We use the abbreviations N (Gauss), F (Frank), C (Clayton), J (Joe) and G (Gumbel). An extensive discussion of bivariate copula families and their properties can be found in Joe (1997) or Nelsen (2006).

Model class	Families	log predictive score
1	N	363.8
2	F	360.6
3	N, F	365.4
4	N, C	356.3
5	N, J	365.5
6	N, F, C	365.1
7	N, F, J	366.4
8	N, F, C, J	366.2
9	N, F, C, J, G	366.4
10	N, F, C, J, G + rotations by 90°, 180° and 270°	366.2

The independence model is outperformed for all choices of family sets. Further, we see no indication of overfitting when more copula families are included, but a loss in predictive performance when some are excluded, in particular for the Gaussian and Frank copula. For this reason, we choose modelclass 9 with Gaussian, Frank, Clayton, Joe and Gumbel copulas for our further analysis since we believe it to offer the best

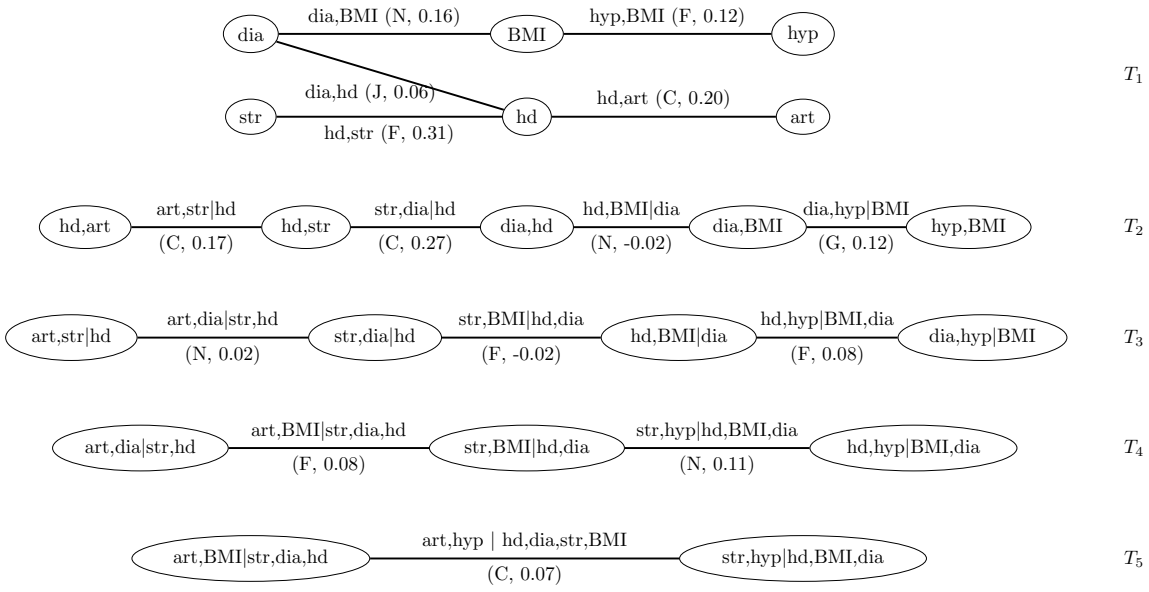
compromise between flexibility, predictive performance and computation tractability. For the whole data, the selected model is the following: We use GLMs for the modeling of marginal response variables (the covariates are available upon request). The dependence between these marginal models is then subsequently described using a discrete-continuous R-vine copula, with R-vine structures and associated pair copulas as shown in Figure 3 for Wave 2. Comparison of the resulting model probabilities with the observed probabilities indicates that the model can accurately describe the observed dependence patterns.

Since the selection procedure of Section 3.2 selects the strongest dependencies (i.e. the dependencies where the corresponding copula terms lead to the biggest improvements in the joint likelihood) first, these are on the first trees of Figure 3. For the first tree T_1, copulas between BMI and diabetes, BMI and hypertension as well as heart disease and stroke are selected for all three waves of observations. This shows that these are the most important dependencies in the data. On the other hand, the copulas on higher trees correspond to weaker conditional dependencies which might even be close to conditional independence.

4.4 Results

Due to restrictions of space for this article, we only present results for Wave 2 in the following. The results for Wave 1 and Wave 3 are available upon request. While modeling and interpreting dependency between purely continuous variables is relatively well understood, this is more challenging in the presence of discrete outcomes. For purely continuous variables, most researchers will look at the theoretical rank correlations and bivariate tail dependencies associated with a copula model, which are usually good summary statistics for the data. In the setup with discrete and continuous outcomes considered here, changes in strength of dependence can be expressed by

Figure 3 The R-vine tree structure, pair copulas and corresponding parameter estimates for the second wave of observations of the six response variables. Here, the pair-copulas are parametrized in terms of the theoretical Kendall's tau values which would result in the purely continuous case. The variables are labeled as follows: hypertension (hyp), diabetes (dia), arthritis (art), heart disease (hd), stroke (str), and body mass index (BMI).



different copula families being selected. In particular, the limiting dependence behavior (for large and small values of the continuous variable, respectively) of conditional distributions is different across copula families. While our inference procedure yields point estimates and standard errors for all model parameters and allows to compute p-values for regression parameters we omit listing these estimates here. Instead, we compute conditional probabilities from our model to better understand the results. For example, we explored the conditional probabilities of each chronic condition given BMI by category of predictors such as age level. Here, conditional probabilities involving marginal covariates are computed as follows: Let $\mathbf{x}_i$ be the vector of covariates

for patient i, $z_1, z_2 \in \mathbb{R}$ and $\mathbf{Y}$ the vector of outcomes. Then

$$P(Y_{hyp.} = 1|BMI = z_1, age \leq z_2) := \sum_{i|x_{i,age} \leq z_2} \frac{P(Y_{hyp.} = 1|BMI = z_1, \mathbf{x}_i)}{\#\{i|x_{i,age} \leq z_2\}},$$

i.e. we average over all relevant covariate vectors in the population. When not conditioning on marginal covariates, we have

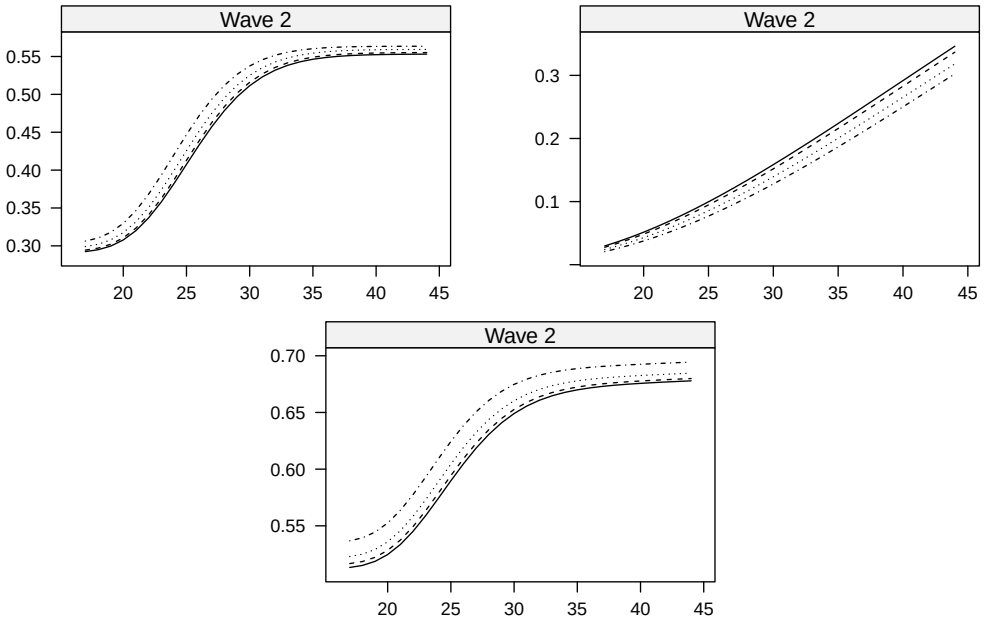
$$P(Y_{hyp.} = 1|BMI = z_1) := \sum_{i=1}^{N} \frac{P(Y_{hyp.} = 1|BMI = z_1, \mathbf{x}_i)}{N},$$

where N is the number of patients. To reduce the computational complexity for producing the plots with the density of BMI given other outcome variables, we show it for an "average" patient in our sample. This means that we have for example

$$f_{BMI}(z|Y_{hyp.} = 1, Y_{art.} = 1) := f_{BMI}(z|Y_{hyp.} = 1, Y_{art.} = 1, \mathbf{x}_{average}),$$

where $\mathbf{x}_{average}$ refers to a female non-smoker with 75.57 years of age at the beginning of the study, and education score of 12.1 and an income score of 17.64. Figure 4 depicts the relationship between a subject's BMI and the conditional probability of hypertension, diabetes and arthritis for the different time periods. The top, middle, and bottom rows represent the patients with hypertension, diabetes and arthritis, respectively. The different lines in each plot correspond to different age groups. The solid line is the mean level for patients of age ≤ 72 at the beginning of the study, dashed for $72 <$ age ≤ 77, dotted for $75 <$ age ≤ 78, and dash-dotted for age > 78.
In Figure 4, we can see that higher probabilities of observing the three diseases (hypertension, diabetes and arthritis) are associated with increasing BMI values. First, the probability of diabetes (upper right panel) is almost linearly increasing with BMI.

Figure 4 Conditional probability of observing hypertension (upper left panel), diabetes (upper right panel) and arthritis (lower panel), respectively, given a certain value for BMI. The solid line is the mean level for patients with ($age \leq 72$) at the beginning of the study (dashed: $72 < age \leq 75$, dotted: $75 < age \leq 78$, dash-dotted: $age > 78$).



This positive association between diabetes and BMI (or obesity) has been reported for all ages (Nguyen et al. 2011) and it is widely accepted that BMI is one of the strongest predictors for diabetes. Therefore, sustained weight loss can bring a reduced risk of diabetes, as studied in Moore et al. (2000). Meanwhile, it is interesting to note that the prevalence of diabetes is slightly lower for the oldest elderly group in our sample, which might be explained by a decline in BMI which is generally observed after about 60 years of age (Elia 2001). A different trend is observed for the prevalence of arthritis with respect to BMI: a family of S-shaped curves in the bottom panel. This confirms a general positive association between BMI and arthritis which has previously been reported in studies for the overall population (Zakkak et al. 2009). However, these studies suggest a stronger increase for the heavily obese (BMI > 40) as compared to the group with $30 <$ BMI < 40 than we find in our sample. This different behavior which we observe might be attributable to a general decline in physical activity in the elderly population, since physical activity is associated both with obesity and with arthritis (Shih et al. 2006). Similar shapes are observable also for hypertension. Though systematic studies are scarce, a general increase of blood pressure with BMI has been previously reported for elderly populations (Masaki et al. 1997).

The different shape of the conditional probability curve of hypertension and arthritis is expressed in the model by different copula families. The shape of the curve is governed by the limits of the conditional distribution (Table 4). While, e.g., the Frank copula has a finite limit for arbitrarily small BMI values, the limit for the Clayton and Gaussian copula is 1. Thus, the probabilities continue to increase.

Figure 5 leverages our joint dependence model to show the complex dependence of the probability of observing arthritis with BMI and other chronic conditions. The upper left panel of Figure 5 shows the probability of arthritis given BMI with the presence/absence of diabetes and the presence/absence of hypertension, thus pro-

Figure 5 Upper left panel: conditional probability of observing arthritis given BMI and other chronic conditions: solid (diabetes, hypertension), dashed (diabetes, no hypertension), dotted (no diabetes, hypertension), dash-dotted (no diabetes, no hypertersion).
Upper right panel: conditional probability of observing diabetes given BMI and other chronic conditions: solid (heart disease, stroke), dashed (heart disease, no stroke), dotted (no heart disease, stroke), dash-dotted (no heart disease, no stroke).
Lower panel: density of BMI given other chronic conditions: solid (hypertension, arthritis), dashed (hypertension, no arthritis), dotted (no hypertension, arthritis), dash-dotted (no hypertension, no arthritis).

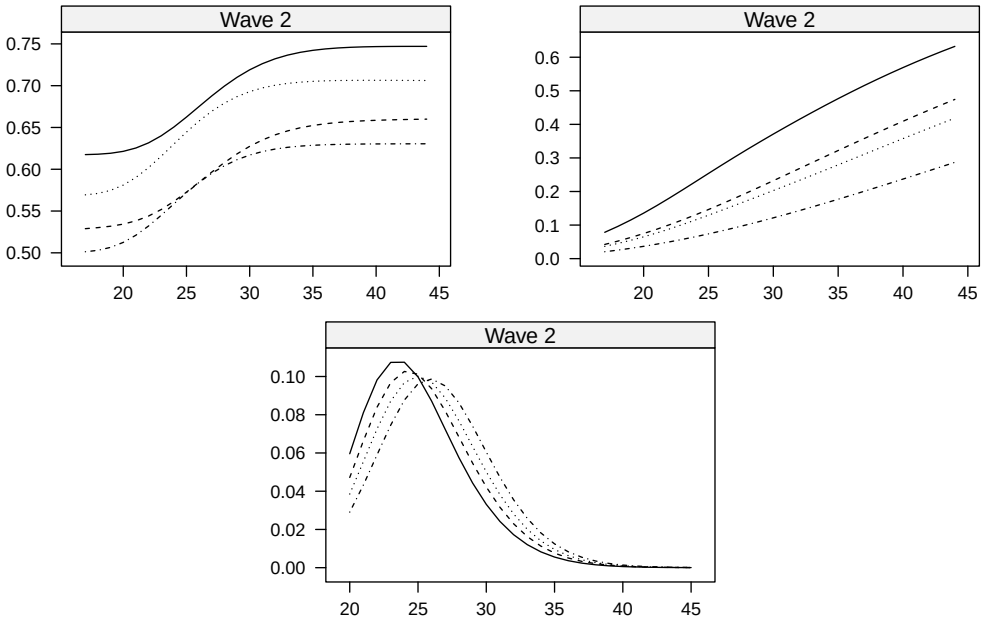


Table 4 The limiting behavior of conditional distribution functions $\partial_2 C(u_1, u_2)$ corresponding to well-known bivariate copula families (see Schepsmeier and Stöber (2014) for details on the pair copula families and parametrization).

Copula Family	$u_2 \to 0$	$u_2 \to 1$
Clayton	1	$(u_1^{-\theta} - 1)^{-(1+1/\theta)}$
Gumbel	1	0
Joe	$(1 - u_1)^{\theta - 1}$	0
Frank	$\frac{e^{\theta}}{e^{\theta u_1}} \frac{e^{\theta u_1} - 1}{e^{\theta} - 1}$	$\frac{e^{\theta u_1} - 1}{e^{\theta} - 1}$
Gauss	1	0

ducing four different plots. This enables us to see the complete picture of arthritis prevalence. The plot indicates positive dependence between arthritis and the other two diseases (diabetes and hypertension). When a subject had both diabetes and hypertension, the probability of having arthritis was higher compared to a subject who suffers from only one or no chronic condition. Likewise, the probability of having arthritis was higher with the obese people (BMI ≥ 30).

The upper right panel of Figure 5 presents the conditional probability of observing diabetes given BMI and two other chronic conditions (heart disease and stroke). The probability of diabetes is not affected strongly by the presence/absence of heart diseases and stroke when the BMI is low, however, as the BMI level increases the chance of arthritis was getting larger depending on the presence of the cardiovascular diseases (CVD). Compared to the case when elderly have either heart disease or stroke, the risk of diabetes jumped by more than 15% for obese patients with both heart disease and stroke, indicating that diabetes is associated with CVD.

The lower panel of Figure 5 finally presents the conditional density of BMI given the absence or presence of two other chronic conditions (heart disease and stroke). We observe that conditioning on different combinations does not only affect the mean of the distribution of BMI but also its variance. We want to note that caution is

needed when interpreting probability plots since the displayed associations do not imply causations. For instance, though we plotted the probability of diabetes given presence/absence of CVD, diabetes is usually considered as the risk factor of CVD in medical literatures. Our plots only serve as a reference to illustrate the multivariate association among the diseases. We reported only selected probabilities due to restrictions of space, additional plots are available upon request.

5 Conclusion

Our aim was to develop a copula model for the joint modeling of discrete and continuous response variables in a regression setup to help understanding comorbidity of the elderly and give new clues about its pathways. Building on the theory of PCCs we developed a flexible model of multivariate association. While competing models for multivariate discrete data are usually fitted using computationally intensive MCMC methods, our model can be rapidly fitted to data sets with several thousand observations. This has been demonstrated using data from the LSOA II, where our model selection heuristic and parameter estimation using maximum likelihood have been applied. Since PCCs allow to combine different copula families, different limiting behavior of conditional probabilities for the presence of diseases given the BMI could also be modeled. This improves the predictive performance of the copula model compared to models where all bivariate families are the same as cross-validation shows.

Despite the success of our proposed method in providing useful information for the health consequences of the elderly, we acknowledge some modeling limitations inherent to the incomplete information in the LSOA II data and the model complexity. One of the complicating aspects of the study with older individuals is that a researcher often confronts with high drop-out rates due to death or some other unknown reasons. Decedents and losses to follow-up were relatively high in the LSOA II data.

Since the reason for dropout is likely to be associated with the elderly's health status, these drop-outs may not be simply ignored. Moreover, in the longitudinal data the responses are recorded over time, at different time points, and these observations within each subject tend to be correlated. In our paper, we did not account for the dropout and the inter-subject correlation due to the model complexity in the proposed copula modeling setting, these can be further considered in our future research.

The response variables were based on a self-report study. Although the study of Kriegsman et al. (1996) implicates that self-reports on chronic diseases are fairly accurate, the use of self-reported diagnoses is another methodological limitation that may have introduced both systematic errors. In particular, Kriegsman et al. (1996) find that self-reports on arthritis were often incorrect. Utilizing clinical interviews or general practitioners information might be better ways to obtain data. Finally, while the inference procedures demonstrated here allow to estimate standard errors for parameter estimates, model uncertainty cannot be addressed. This could be done in a computationally more intense RJMCMC framework (c.f. Czado et al. (2013)).

Since the multi-dimensional mixed type of responses often appear in both cross-sectional and longitudinal data, the proposed method can be applied to other applications in similar settings. (e.g, our approach can be adapted by a clinician who desires to estimate the patient's current status in multiple dimensions.) We hope that our integrated analysis of the relationships among chronic conditions in the older people will improve geriatric assessment and may be used in health service evaluation.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula construction of multiple dependence. *Insurance: Mathematics and Economics 44*, 182–198.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transaction*

on Automatic Control 19(6), 716–723.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Bedford, T. and R. Cooke (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence 32*, 245–268.

Bedford, T. and R. Cooke (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics 30*, 1031–1068.

Brechmann, E., C. Czado, and K. Aas (2012). Truncated regular vines in high dimensions with applications to financial data. *Canadian Journal of Statistics 40*(1), 68–85.

Czado, C., E. Brechmann, and L. Gruber (2013). Selection of vine copulas. In *Copulae in Mathematical and Quantitative Finance*. Springer.

Czado, C., U. Schepsmeier, and A. Min (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling 12*, 229–255.

D. Hoff, P. (2007, 06). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat. 1*(1), 265–283.

Danaher, P. J. and M. S. Smith (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science 30*(1), 4–21.

Dißmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis 59*, 52–69.

Dobra, A. and A. Lenkoski (2011, 06). Copula gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat. 5*(2A), 969–993.

Elia, M. (2001). Obesity in the elderly. *Obesity Research 9*, 244–248.

Fleischer, N., R. A. Diez, M. Alazraqui, H. Spinelli, and F. De Maio (2011). Socioeconomic gradients in chronic disease risk factors in middle-income countries: evidence of effect modification by urbanicity in argentina. *Am J Public Health 101*, 294–301.

Flicker, L., K. A. McCaul, G. J. Hankey, K. Jamrozik, W. J. Brown, J. E. Byles, and O. P. Almeida (2010). Body mass index and survival in men and women aged 70 to 75. *Journal of the American Geriatrics Society 58*(2), 234–241.

Fonseca, M. d. J. M. d., V. L. Andreozzi, E. Faerstein, D. Chor, and M. S. Carvalho (2008, 02). Alternatives in modeling of body mass index as a continuous response variable and relevance of residual analysis . *Cadernos de Sa'ude P'ublica 24*, 473–478.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika 82*, 543–552.

Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Astin Bulletin 37*(2), 475–515.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal*

Statistical Society. Series B (Methodological) 46(2), 149–192.

He, J., H. Li, A. Edmondson, D. Rader, and M. Li (2012). A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics 13*, 497–508.

Hong, H. G., Y. Yue, and P. Gosh (2013). Bayesian estimation of long-term health consequences of obese and normal-weight elderly. *preprint*.

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

Joe, H., H. Li, and A. K. Nikoloulopoulos (2010, January). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis 101*(1), 252–270.

Joe, H. and J. J. Xu (1996). The estimation method of inference functions for margins for multivariate models. *UBC, Dept. of Statistics, Technical Report 166*.

Kriegsman, D., B. Penninx, J. van Eijk, A. Boeke, and D. Deeg (1996). Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. a study on the accuracy of patients' self-reports and on determinants of inaccuracy. *Journal of Clinical Epidemiology 49*, 1407–1417.

Leon, A. R. d. and W. B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine 30*, 175–185.

Li, J. and W. K. Wong (2011). Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness. *Biostatistics 12*(1), 143–155.

Masaki, K. H., J. D. Curb, D. Chiu, H. Petrovitch, and B. L. Rodriguez (1997). Association of body mass index with blood pressure in elderly japanese american

men: The honolulu heart program. *Hypertension 29*(2), 673–677.

Masarotto, G. and C. Varin (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics 6*, 1517–1549.

Min, A. and C. Czado (2011). Bayesian model selection for D-vine pair-copula constructions. *Canadian Journal of Statistics 39*(2), 239–258.

Moore, L., A. Visioni, P. Wilson, R. D'Agostino, W. Finkle, and R. Ellison (2000). Can sustained weight loss in overweight individuals reduce the risk of diabetes mellitus? *Epidemiology 11*(3), 269–273.

Murray, J., D. Dunson, L. Carin, and J. Lucas (2013). Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association 108*, 656–665.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

Nešlehová, J. (2007, March). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis 98*(3), 544–567.

Nguyen, N., X. Nguyen, J. Lane, and P. Wang (2011). Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 1999-2006. *Obesity Surgery 21*, 351–355.

Nikoloulopoulos, A. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference 143*, 1923–1937.

Nikoloulopoulos, A. K. and D. Karlis (2006). Modeling multivariate count data. In A. Rizzi and M. Vichi (Eds.), *Proceedings in Computational Statistics*, pp. 599–606. The International Association for Statistical Computing.

Nikoloulopoulos, A. K. and D. Karlis (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference 139*(11), 3878–3890.

Panagiotelis, A., C. Czado, and H. Joe (2012). Regular vine distributions for discrete data. *Journal of the American Statistical Association 105*(499), 1063–1072.

Pitt, M., D. Chan, and R. Kohn (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika 93*(3), 537–554.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal 36*, 1389–1401.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schepsmeier, U. and J. Stöber (2014). Derivatives and Fisher information of bivariate copulas. *Statistical Papers 55*, 525–542.

Shen, C. and L. Weissfeld (2006). A copula model for repeated measurements with non-ignorable non-monotone missing outcome. *Statistics in Medicine 30*, 2427–2440.

Shih, M., J. Hootman, J. Kruger, and C. Helmick (2006). Physical activity in men and women with arthritis: National health interview survey, 2002. *American Journal of Preventive Medicine 30*, 385–393.

Singh, P. N., E. Haddad, S. Tonstad, and G. E. Fraser (2011). Does excess body fat maintained after the seventh decade decrease life expectancy? *Journal of the American Geriatrics Society 59*(6), 1003–1011.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris 8*, 229–231.

Smith, M. S. and M. A. Khaled (2012). Estimation of copula models with discrete margins via bayesian data augmentation. *Journal of the American Statistical Association 107*(497), 290–303.

Smith, M. S., A. Min, C. Almeida, and C. Czado (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association 105*(492), 1467–1479.

Song, P. X.-K., M. Li, and Y. Yuan (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics 65*(1), 60–68.

Stöber, J., H. Joe, and C. Czado (2013). Simplified pair copula constructions — limitations and extensions. *Journal of Multivariate Analysis 119*, 101–118.

Stöber, J. and U. Schepsmeier (2013). Estimating standard errors in regular vine copula models. *Computational Statistics 28*, 2679–2707.

Zakkak, J., D. Wilson, and J. Lanier (2009). The association between body mass index and arthritis among US adults: CDC's surveillance case definition. *Preventing Chronic Disease 6*(2), 14:1–14:11.