Pair Copula Constructions for Discrete Data

Anastasios Panagiotelis, Claudia Czado and Harry Joe

February, 2011

Abstract

Copulas can be used to construct multivariate models with a wide range of dependence features. Existing approaches to copula modeling for m-dimensional discrete data suffer computational disadvantages as m increases. For copulas with a closed form, the number of terms that must be computed to evaluate the likelihood grows exponentially with m, while for popular elliptical copulas, integration by computationally intensive numerical techniques is required. Although, composite likelihood methods provide a faster alternative, they lack statistical efficiency.

Our contribution is to extend the principles of vine Pair Copula Constructions (PCCs) to discrete margins. Our construction is flexible, even compared to elliptical copulas, since combining different pair copulas as building blocks leads to different joint probabilities. A second major advantage of our approach is that computing the probability mass function requires only $2m(m-1)$ evaluations of bivariate copula functions. Consequently, maximum likelihood estimation of the marginal and copula parameters, whether in a 2-step or joint procedure, is computationally feasible even in high dimensions.

We demonstrate the high quality of maximum likelihood estimates with bootstrapped confidence intervals under a simulated setting. We also illustrate the inferential potential of our model in two real data applications. Further, we make an attempt at addressing model selection issues and outline interesting new directions for future research in the modeling of multivariate discrete data.

Keywords: Generalized Poisson regression, Inference functions for margins, Longitudinal data, Model selection, Ordered probit regression,Vines.

Dr. Anastasios Panagiotelis is a Humboldt Fellow with Professor Claudia Czado at Zentrum Mathematik at the Technische Universität München. Harry Joe is a Professor at the Department of Statistics at the University of British Columbia. Address for correspondence is: Anastasios Panagiotelis, Technische Universität München, Zentrum Mathematik, Lehrstuhl für Mathematische Statistik, Boltzmannstrasse 3, 85748 Garching b. München, Germany. Email: panagiot@ma.tum.de

1 Introduction

The statistical modeling of multivariate discrete data such as binary, categorical and count data has many interesting applications in fields as diverse as finance, biostatistics, econometrics, psychometrics and marketing to name but a few. While the primary aim of such models is to describe dependence, they can also improve the quality of estimates of the marginal parameters. While multivariate analogues of popular models for discrete data such as the multivariate probit and multivariate Poisson have been developed, (see Johnson et al. (1997) and references therein), they can lack generality and flexibility in the type of dependence they capture. Copulas provide a means to construct multivariate discrete distributions from virtually any combination of marginal models and have the potential to attain a wide range of dependence characteristics.

The literature on copula modeling of discrete data covers a wide range of copula families and estimation methods. For copula functions with a closed form, maximum likelihood estimation is straightforward. The probability mass function for m-dimensional data can be computed by simply taking 2^m finite differences of the copula function. This approach is computationally intensive, and becomes infeasible for high-dimensional problems. Consequently, this approach is most frequently used with copulas that are fast to compute. For example, Li and Wong (2011) use a variation of the Farlie-Gumbel-Morgenstern (FGM) family of copulas, Nikoloulopoulos and Karlis (2008) use a mixture of max-id copulas and Nikoloulopoulos and Karlis (2009) use a copula based on a finite normal mixture. Unfortunately these copulas achieve limited dependence; the mixture of max-id cannot capture negative dependence, while the FGM and finite normal mixture copulas capture weak dependence when compared to more flexible elliptical and vine copulas.

There have also been significant efforts to apply elliptical copulas, and in particular the Gaussian copula, to discrete data. This is due to the many attractive properties of elliptical copulas, they capture a wide range of dependence that can be both positive and negative and they are closed under marginalization. Furthermore, many popular models for multivariate discrete data used in econometrics and psychometrics, such as the multivariate probit and ordered probit, can also be expressed as a model involving an elliptical copula. As many elliptical copulas, including the Gaussian copula, cannot be written down in closed form, maximum likelihood estimation through taking finite differences is not a feasible option. Instead, likelihood-based estimation requires integration of the density function of the Gaussian distribution over a rectangle. This integral can be evaluated either numerically or by simulation techniques (see for example Genz (1992) and Joe (1995)). In a similar vein, Bayesian methods can also be used to estimate models based on elliptical copulas (see Pitt et al. (2006) for the Gaussian copula and Smith et al. (2011) for the skew t copula). In general, both Frequentist and Bayesian techniques are computationally intensive, and may not scale easily to higher dimensions. Another alternative, used for example in Zhao and Joe (2005), is to base estimation on the composite likelihood (see Varin et al. (2011) for an general overview of composite likelihood estimation). While such methods are faster, they are not as statistically efficient as maximum likelihood estimation.

Our contribution is to develop a discrete analogue to vine Pair Copula Constructions (PCCs). The advantages of our approach are twofold. First, vine PCCs provide a highly flexible framework for constructing copulas exhibiting a wide range of dependence characteristics, even when compared to elliptical copulas. This flexibility arises since any combination of bivariate copulas can be used to construct vine PCC models (see Czado (2010) for a review). Despite differences with the continuous case, our vine PCCs for discrete margins yield models with joint probabilities that vary when different classes of pair copulas are used, suggesting that flexibility is also an advantage of our construction. Our second major contribution is that computation of the probability mass function for a discrete vine PCC only requires the evaluation of $2m(m-1)$ bivariate copula functions, compared to 2^m multivariate copulas for the finite differences approach. As such, maximum likelihood estimation is feasible even for higher dimensions. Maximum likelihood estimation also compares favorably with the methods used to estimate models based on elliptical copulas, which are either less efficient in a computational or statistical sense. Furthermore, as we shall see in Section 4, model selection techniques can be used to find more parsimonious D-vine models. Given the computational advantages and ability to exploit sparsity in a vine framework, we believe our approach has high potential to be used in truly high-dimensional settings.

The following is a brief summary of the remainder of the paper. Section 2 gives necessary background on copula modeling in general with some specific remarks made on copula modeling for discrete margins. We also review the fast developing literature on vine PCCs for continuous margins, introducing key results and concepts that illuminate and clarify the development of our own model. Section 3 introduces our discrete vine decomposition and discusses the advantages of our approach compared to alternatives. Section 4 outlines maximum likelihood estimation for the model introduced in Section 3 using both a 2-step inference functions for margins (IFM approach) as well as joint estimation. The results of an extensive simulation study are also summarized. Finally, two applications are covered in Section 5; the first involves finance micro-structure and is an analysis of count data from order books with generalized Poisson regression models used for the margins, while the second, coming from biostatistics, is an intra-day analysis of headache severity for a pooled dataset of 135 patients. In Section 6 we review the major contributions of the paper and point to potential areas of future research.

2 Background to Copulas and vine Pair Copula Constructions (PCCs)

In the following section, we review some fundamental concepts, both on copulas generally, and on vine PCCs in the continuous case more specifically. The aim here is twofold. First, we highlight some important distinctions in modeling discrete and continuous data via a copula approach. Second, we provide background for the introduction of discrete D-vine PCC in Section 3. Throughout the paper we will use the following conventions for notation. Probability mass functions (pmf) and cumulative distribution functions (cdf) will be denoted f_Z and F_Z respectively with subscript denoting a random variable or vector. If $\mathbf{Z} := (Z_1, Z_2, \dots, Z_m)'$ is a vector, marginal and conditional distributions are also abbreviated as $F_j := F_{Z_j}$ or $f_j := f_{Z_j}$, $F_{jk} := F_{Z_j,Z_k}$, $F_{j|k} := F_{Z_j|Z_k}$ etc. No subscript on f or F indicates the full multivariate distribution.

2.1 Sklar's Theorem

Over the last 20 years copula models have been extensively applied to a variety of fields. These include, but are not limited to, finance (Embrechts et al. (2003), Patton (2004)), marketing (Danaher and Smith (2011)), health care utilization (Cameron et al. (2004)) and survival analysis (Hougaard (2000)). Introductory material on copulas can be found in Joe (1997), Nelsen (2006) and references therein.

The theorem of Sklar (1959) underpins the literature on copula modeling. Let $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_m)'$ be an m-dimensional random vector with realization $\mathbf{y} := (y_1, y_2, \ldots, y_m)'$, marginal distribution functions $F_j(y_j)$ for $j = 1, 2, \ldots, m$, and joint distribution function $F(y_1, \ldots, y_m)$. Throughout the paper it should be clear from the context, whether the response $\mathbf{Y}$ is discrete or continuous. Sklar's Theorem states

$$F(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)), \tag{2.1}$$

where C is known as a copula function. For continuous $\mathbf{Y}$, the copula function is unique, and is itself an m-dimensional distribution function over the unit hypercube. For discrete random variables, the copula function is unique only over the domain of $\mathbf{Y}$.

In the continuous case, an expression for the multivariate density $f(y_1, \ldots, y_m)$ can be obtained by differentiating both sides of Equation (2.1), yielding

$$f(y_1, \ldots, y_m) = c(F_1(y_1), \ldots, F_m(y_m))f_1(y_1) \ldots f_m(y_m), \tag{2.2}$$

where $f_j(y_j)$ is the marginal density of the j^{th} margin and $c(.)$, known as the copula density, is the the copula function differentiated with respect to each of its arguments. Sklar's Theorem elegantly provides a framework whereby margins can be modeled generally, while the dependence structure is described by the copula. A wide range of both parametric and non-parametric approaches to modeling the copula function have been developed.

Although the copula function is not unique for discrete margins, parametric copulas may still be used to model the dependence between discrete data. For further investigation of this issue we refer the reader to Genest and Nešlehová (2007) who provide some evidence that discrete data inherit dependence properties from a parametric copula in a similar way to the continuous case. In contrast with the continuous case, there are, broadly speaking, two ways to evaluate the probability mass function for discrete data. In general, the pmf can be evaluated by taking differences of the copula function. Assuming for simplicity, and without loss of generality that $\mathbf{Y} \in \mathbb{N}^m$ (where $\mathbb{N}$ is the set of natural numbers), the probability mass function of $\mathbf{Y}$ is given by

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{i_1=0,1} \cdots \sum_{i_m=0,1} (-1)^{i_1+\ldots+i_m} \Pr(Y_1 \leq y_1 - i_1, \ldots, Y_m \leq y_m - i_m) \\ &= \sum_{i_1=0,1} \cdots \sum_{i_m=0,1} (-1)^{i_1+\ldots+i_m} C(F_1(y_1 - i_1), \ldots, F_m(y_m - i_m)) \,. \qquad (2.3) \end{aligned}$$

Computing the right hand side of Equation (2.3) requires 2^m evaluations of a multivariate copula function.

Alternately, the pmf for some copula functions that do not have a closed form can be evaluated by integration over a rectangle. For example, letting $\psi_j^+ := \Phi^{-1}(\Pr(Y_j \leq y_j))$, where $\Phi^{-1}(.)$ denotes the inverse cdf of a univariate standard normal distribution, the popular Gaussian copula can be expressed as

$$C(y_1, \ldots, y_m) = \Phi_m(\psi_1^+, \ldots, \psi_m^+; \Gamma) = \int_{-\infty}^{\psi_1^+} \cdots \int_{-\infty}^{\psi_m^+} \phi_m(\psi_1, \ldots, \psi_m; \Gamma) d\psi_1 \ldots d\psi_m \,, \qquad (2.4)$$

where $\Phi_m(., \Gamma)$ and $\phi_m(., \Gamma)$ respectively denote the cdf and probability density function of an m-dimensional normal distribution with mean 0 and variance matrix given by the correlation matrix Γ. For the Gaussian copula with discrete margins, the probability mass function can be expressed as follows

$$\Pr(Y_1 = y_1, \ldots, Y_m \leq y_m) = \int_{\psi_1^-}^{\psi_1^+} \cdots \int_{\psi_m^-}^{\psi_m^+} \phi_m(\psi_1, \ldots, \psi_m; \Gamma) d\psi_1 \ldots d\psi_m \,, \qquad (2.5)$$

where $\psi_j^- := \Phi^{-1}(\Pr(Y_j = y_j - 1))$. Despite notable advances in evaluating such integrals, some of which are mentioned in Section 1, this remains a highly challenging computational problem especially for higher dimensions. Furthermore, such techniques do not generalise to other copulas such as vines. The D-vine decomposition we propose in Section 3 is a copula based framework that greatly reduces the computational cost of evaluating the pmf, but is also highly flexible, meaning that a large range of dependence characteristics can be modeled.

2.2 PCCs in the continuous case

A general problem in copula modeling is that although the class of bivariate copula families is quite rich, there is a paucity of copulas that generalize to higher dimensions. Vine Pair Copula Constructions (PCCs), developed over a series of papers including Joe (1996) Bedford and Cooke (2001) and Bedford and Cooke (2002), and first used in an inferential context by Kurowicka and Cooke (2006) and Aas et al. (2009), provide a principle for constructing multivariate copulas as the product of bivariate pair copula building blocks. Their flexibility can capture a wide range of symmetric or asymmetric tail dependence (Joe et al. (2010)), they nest popular elliptical copulas (Haff et al. (2010)) and they have been shown to be highly competitive in the modeling of empirical data from finance (Fischer et al. (2009)). Recent developments and applications of PCC based models for continuous data can be found in Kurowicka and Joe (2011). We now briefly review some key concepts for vine PCCs in the continuous case before introducing discrete vine PCCs in Section 3.

For continuous $\mathbf{Y}$, a PCC is derived by starting with the following decomposition

$$f(y_1, \ldots, y_m) = f_{1|2\cdots m}(y_1|y_2, \ldots, y_m)\, f_{2|3\cdots m}(y_2|y_3, \ldots, y_m) \cdots f_m(y_m)\,. \tag{2.6}$$

Each term on the right hand side is of the general form $f_{Y_j|\mathbf{V}}(y_j|\mathbf{v})$ where y_j is a scalar element of $\mathbf{y}$ and $\mathbf{V}$ is a subset of $\mathbf{Y}$. Using Sklar's theorem, we can show that any such univariate conditional density can be decomposed into the product of a bivariate copula density and another univariate conditional density with the conditioning set reduced by one element. Letting V_h be any scalar element of $\mathbf{V}$ and $\mathbf{V}_{\backslash h}$ its complement, with Y_j not an element of $\mathbf{V}$,

$$f_{Y_j|\mathbf{V}} = \frac{f_{Y_j,V_h|\mathbf{V}_{\backslash h}}}{f_{V_h|\mathbf{V}_{\backslash h}}}\,. \tag{2.7}$$

Applying Equation (2.2) for conditional densities to the numerator gives the following

$$\begin{aligned} f_{Y_j|\mathbf{V}} &= \frac{c_{Y_j,V_h|\mathbf{V}_{\backslash h}}(F_{Y_j|\mathbf{V}_{\backslash h}}, F_{V_h|\mathbf{V}_{\backslash h}}) f_{Y_j|\mathbf{V}_{\backslash h}} f_{V_h|\mathbf{V}_{\backslash h}}}{f_{V_h|\mathbf{V}_{\backslash h}}} \\ &= c_{Y_j,V_h|\mathbf{V}_{\backslash h}}(F_{Y_j|\mathbf{V}_{\backslash h}}, F_{V_h|\mathbf{V}_{\backslash h}}) f_{Y_j|\mathbf{V}_{\backslash h}} \end{aligned} \tag{2.8}$$

where $c_{Y_j,V_h|\mathbf{V}_{\backslash h}}$ denotes the pair copula density describing the dependence between Y_j and V_h conditional on $\mathbf{V}_{\backslash h} = \mathbf{v}_{\backslash h}$. This expression can be applied recursively to each of the terms in Equation (2.6) until the multivariate density is decomposed into the product of $m(m-1)/2$ bivariate copulas, which is itself a multivariate copula, and the product of the marginal densities. As V_h can be any element of $\mathbf{V}$ at Equation (2.7) there are several ways that a density can be decomposed in this manner. However, for statistical modelling, it is desirable to find a set of decompositions whereby the arguments of the copula densities can be computed using pair copulas that are defined elsewhere in the PCC. The different ways that a density can be decomposed to satisfy this condition were organized and summarized by Bedford and Cooke (2001) and Bedford and Cooke (2002) in terms of graphical models called vines.

Before discussing vines in more detail in Section 2.3, we briefly digress to show how PCCs can be used to construct flexible multivariate models using the following 3-dimensional vine pair copula construction

$$f(y_1, y_2, y_3) = c_{13|2}(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2))c_{12}(F_1(y_1), F_2(y_2))c_{23}(F_1(y_2), F_2(y_3)) \,, \qquad (2.9)$$

as an illustrative example. This decomposition can motivate a statistical model if we assume that the conditional copulas depend on the conditioning set only through their arguments. For example, in Equation (2.9), the functional form of the copula $c_{13|2}(.,.)$ would not depend on y_2. Typically parametric bivariate copulas such as the MTCJ/Clayton[1], Gumbel, Gaussian, Student t or Frank Copulas will be chosen to model the pair copulas. Although, strictly speaking, notation of the form $c_{13|2}(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2); \theta_{13|2})$, where $\theta_{13|2}$ are the copula parameters, should be used, the copula parameters are dropped for convenience unless it is confusing to do so. Also, while we only consider one-parameter families in this paper, $\theta_{13|2}$ can be vector-valued with no loss of generality. As the number of parametric bivariate copula families is quite large and as any combination of bivariate copulas may be used in a PCC, this modeling approach is highly flexible.

The arguments of the pair copulas are conditional distribution functions and can be

[1]Although the MTCJ copula should be attributed to Mardia, Takashi, Cook and Johnson (see Joe et al. (2010) for a detailed discussion), the bivariate version is more commonly referred to as the Clayton copula. We use the name MTCJ/Clayton throughout the paper as a compromise between recognising authors who made the biggest contributions in discovering the copula and avoiding confusion.

evaluated using the following expression given by Joe (1996)

$$F_{Y_j|V_h,\mathbf{v}_{\setminus h}}(y_j|v_h,\mathbf{v}_{\setminus h}) = \frac{\partial C_{Y_j,V_h|\mathbf{V}_{\setminus h}}(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j|\mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h|\mathbf{v}_{\setminus h}))}{\partial F_{V_h|\mathbf{V}_{\setminus h}}(v_h|\mathbf{v}_{\setminus h})}. \tag{2.10}$$

For most parametric copulas it is straightforward to evaluate this derivative analytically and algorithms that recursively compute the density of a PCC can consequently be constructed (see Aas et al. (2009)).

2.3 Vines

We now turn our attention to explaining vines using the specific example of a 5-dimensional D-vine shown in Figure 1. The D-vine is chosen, since the algorithms outlined in Section 3.3 are based on a discrete version of the D-vine, and because the D-vine has certain advantages in applications where some intuitive ordering of the margins can be made such as longitudinal models for intra-day data. For further details on other classes of vines, we direct the reader to Czado et al. (2010) for C-vines and Dissmann (2010) for R-vines.

——Figure 1 about here——

A vine is characterized by $m-1$ trees denoted T_j for $j = 1,\ldots,m-1$. The j^{th} tree is made up of nodes, denoted N_j and edges which join these nodes, denoted E_j. In the first tree of a D-vine, the margins are ordered, and the edges simply join adjacent nodes yielding $E_1 := \{12, 23, 34, 45\}$. The edges on the first tree become the nodes on the second tree and in general $N_{j+1} := E_j$. The edges of trees $T_2,\ldots,T_{m-1}$ also connect adjacent nodes. Any element shared by two nodes will be in the conditioning set of the edge joining them. For example, the edge joining node 12 and 23 is 13|2, while the edge joining 24|3 and 35|4, will be 25|34. The pair copulas that make up the corresponding PCC are simply indicated by the edges of the entire vine $\{E_1, E_2, \ldots, E_{m-1}\}$, so that the density for a 5-dimensional PCC is given by

$$f(y_1,\ldots,y_5) = c_{12}c_{23}c_{34}c_{45}c_{13|2}c_{24|3}c_{35|4}c_{14|23}c_{25|34}c_{15|234}f_1f_2f_3f_4f_5\,, \tag{2.11}$$

where the arguments of the pair copulas and density functions have been dropped for ease of notation (for example, $c_{14|23}$ is short for $c_{14|23}(F_{1|23}(y_1|y_2, y_3), F_{4|23}(y_4|y_2, y_3)))$. For the D-vine, selecting a different initial ordering of the margins allows different dependence features to be uncovered.

In Section 5 we consider two applications where some intuitive ordering of the margins is available. For applications where no such sensible ordering exists, the margins can be ordered so that pairwise Kendall's τ in the first tree are maximized (Aas et al. (2009)). Although choosing a specific ordering does impose some restriction on our model as D-vines with a different ordering are non-nested, a large degree of flexibility is still available as different families can be chosen for each pair copula. Furthermore, in the continuous case, popular multivariate copulas such as the Gaussian and Student-t copulas are always nested within D-vines for any ordering of the variables. For these reasons we restrict our attention to D-vines in this paper while acknowledging that methods for choosing the correct ordering of margins in a D-vine or even selecting more general vine structures are an interesting area for future research. Some inroads into this problem have already been made in the continuous case by Dissmann (2010) and Brechmann et al. (2010), and it would be interesting to see whether these methods can be extended to the discrete margin case.

3 Discrete D-Vine

In the following section we introduce vine PCCs for discrete margins. First, we outline the discrete analogues to some important equations introduced in Section 2.2. Second, a simple 3-dimensional vine is decomposed in full detail for illustrative purposes. Third, we discuss the advantages of our D-vine decomposition, namely that it gives flexible models with parameters that can be estimated in a computationally and statistically efficient manner.

3.1 Discrete PCCs

Our aim is to decompose a general multivariate pmf into bivariate pair copula building blocks. Using similar notation to Section 2, we can decompose a probability mass function as follows

$$\Pr(Y_1 = y_1, \ldots, Y_m = y_m) = \Pr(Y_1 = y_1 | Y_2 = y_2, \ldots, Y_m = y_m) \times \\ \Pr(Y_2 = y_2 | Y_3 = y_3, \ldots, Y_m = y_m) \times \cdots \times \Pr(Y_m = y_m)\,. \tag{3.1}$$

This expression can be thought of as the discrete analogue to the decomposition at Equation (2.6). Here, we have an expression where each term is of the form $\Pr(Y_j = y_j | \mathbf{V} = \mathbf{v})$ where y_j is a scalar element of $\mathbf{y}$ and $\mathbf{v}$ is a subset of $\mathbf{y}$. In a similar fashion to the continuous case we choose a single element of $\mathbf{v}$ and obtain the following discrete analogue to Equation (2.7).

$$\Pr(Y_j = y_j | \mathbf{V} = \mathbf{v}) = \frac{\Pr(Y_j = y_j, V_h = v_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h})}{\Pr(V_h = v_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h})}\,. \tag{3.2}$$

Assuming $\mathbf{Y} \in \mathbb{N}^m$ for simplicity, the pmfs used in the numerator can be expressed as

$$= \frac{\sum\limits_{i_j=0,1} \sum\limits_{i_h=0,1} (-1)^{i_j+i_h} \Pr(Y_j \leq y_j - i_j, V_h \leq v_h - i_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h})}{\Pr(V_h = v_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h})}\,. \tag{3.3}$$

The bivariate conditional probability in the numerator can be expressed in terms of a copula giving

$$= \frac{\sum\limits_{i_j=0,1} \sum\limits_{i_h=0,1} (-1)^{i_j+i_h} C_{Y_j,V_h|\mathbf{V}_{\setminus h}}(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j - i_j | \mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h - i_h | \mathbf{v}_{\setminus h}))}{\Pr(V_h = v_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h})}\,, \tag{3.4}$$

where $F_{A|B}(a|b)$ is used as generic notation for the distribution function of $\Pr(A \leq a | B = b)$. Equation (3.4) provides a discrete analogue to Equation (2.8) in that it can be applied recursively to Equation (3.1) to decompose a multivariate pmf into smaller bivariate copula building blocks. Finally, the arguments of the copula functions in Equation (3.4) are evaluated using the following

$$F_{Y_j|V_h,\mathbf{V}_{\setminus h}}(y_j | v_h, \mathbf{v}_{\setminus h}) = \Big[C_{Y_j,V_h|\mathbf{V}_{\setminus h}} \Big(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j | \mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h | \mathbf{v}_{\setminus h}) \Big) - \\ C_{Y_j,V_h|\mathbf{V}_{\setminus h}} \Big(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j | \mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h - 1 | \mathbf{v}_{\setminus h}) \Big) \Big] \Big/ \Pr(V_h = v_h | \mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h}) \tag{3.5}$$

which is a discrete analogue to Equation (2.10). For illustration we now outline such a decomposition in detail for the 3-dimensional case.

3.2 A 3-dimensional illustration

Using the same notation as the previous subsection, and setting $m = 3$ we now decompose the probability mass function of the joint density,

$$\begin{aligned}\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) &= \Pr(Y_1 = y_1 |, Y_2 = y_2, Y_3 = y_3) \times \\ & \quad \Pr(Y_3 = y_3 |, Y_2 = y_2) \times \Pr(Y_2 = y_2)\,. \end{aligned} \tag{3.6}$$

Applying Equation (3.4) to the first term on the right hand side with $Y_j = Y_1$, $V_h = Y_3$, $V_{-\backslash h} = Y_2$ gives

$$\Pr(Y_1 = y_1 | Y_2 = y_2, Y_3 = y_3) = \frac{\sum_{i_1=0,1} \sum_{i_3=0,1} (-1)^{i_1+i_3} C_{13|2}(F_{1|2}(y_1 - i_1|y_2), F_{3|2}(y_3 - i_3|y_2))}{\Pr(Y_3 = y_3|Y_2 = y_2)}\,. \tag{3.7}$$

Utilizing Equation (3.5), the first argument of the copula function in the numerator of Equation (3.7) is given by

$$F_{1|2}(y_1 - i_1|y_2) = \frac{C_{12}(F_1(y_1 - i_1), F_2(y_2)) - C_{12}(F_1(y_1 - i_1), F_2(y_2 - 1))}{\Pr(Y_2 = y_2)}\,, \tag{3.8}$$

while the second argument can be expressed as

$$F_{3|2}(y_3 - i_3|y_2)) = \frac{C_{23}(F_2(y_2), F_3(y_3 - i_3)) - C_{23}(F_2(y_2 - 1), F_3(y_3 - i_3))}{\Pr(Y_2 = y_2)}\,. \tag{3.9}$$

Noting that the denominator of Equation (3.7) cancels with the second term on the right hand side of Equation (3.6), the full expression for the probability mass function 3-dimensional discrete D-vine is given by

$$\begin{aligned}&\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \\ &\Big\{ \sum_{i_1=0,1} \sum_{i_3=0,1} (-1)^{i_1+i_3} C_{13|2}\Big(\frac{C_{12}(F_1(y_1 - i_1), F_2(y_2)) - C_{12}(F_1(y_1 - i_1), F_2(y_2 - 1))}{F_2(y_2) - F_2(y_2 - 1)}, \\ &\frac{C_{23}(F_2(y_2), F_3(y_3 - i_3)) - C_{23}(F_2(y_2 - 1), F_3(y_3 - i_3))}{F_2(y_2) - F_2(y_2 - 1)}\Big)\Big\} \left[F_2(y_2) - F_2(y_2 - 1)\right]\,.\end{aligned} \tag{3.10}$$

Although this decomposition differs from that in the continuous case, they both decompose a multivariate density or probability mass function into smaller bivariate pair copula building blocks. Note that in contrast to the continuous case, here copula and marginal distribution functions must be computed, as opposed to copula and marginal densities.

3.3 Discrete D-Vine

As was the case for continuous margins, vines can be used to describe and summarize different discrete PCCs. In Appendix A we outline an algorithm for computing the probability mass function of a D-vine in general dimension. It is evident both from this algorithm, and from the 3-dimensional example in Section 3.2 that each bivariate pair copula only needs to be evaluated 4 times, specifically $C_{Y_j,V_h|\mathbf{V}_{\setminus h}}$ must be evaluated at $(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j|\mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h|\mathbf{v}_{\setminus h}))$, $(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j-1|\mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h|\mathbf{v}_{\setminus h}))$,$(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j|\mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h-1|\mathbf{v}_{\setminus h}))$ and $(F_{Y_j|\mathbf{V}_{\setminus h}}(y_j-1|\mathbf{v}_{\setminus h}), F_{V_h|\mathbf{V}_{\setminus h}}(v_h-1|\mathbf{v}_{\setminus h}))$. As any vine PCC is composed of only $m(m-1)/2$ pair copulas, evaluation of the probability mass function requires only $2m(m-1)$ evaluations of bivariate copula functions. We reiterate that vine PCCs have great potential in high-dimensional settings since the computational burden of evaluating the pmf grows quadratically with m as opposed to exponentially as is the case for the finite differences approach.

A major advantage of D-vine PCCs in general is that a wide variety of dependence structures can be modeled by selecting different copula families as building blocks. Although many dependence concepts such as tail dependence degenerate in the discrete case, selecting different copula families in a discrete D-vine has a substantial impact on the joint probabilities of the multivariate distribution. To demonstrate this point, we consider a 5-dimensional D-vine with Bernoulli margins. We consider 4 cases summarized in Table 1, and for each case set all pair copulas to Gaussian, all pair copulas to MTCJ/Clayton and all pair copulas to Gumbel, giving 12 models overall. Within each case, the marginal probabilities and dependence as measured by Kendall's τ are the same and consequently any difference in the joint probabilities only reflect the use of different copula families as building blocks. In all cases, Kendall's τs are highest in the first tree, and become lower in subsequent trees, as this pattern is consistent with what we observe in many datasets including those is Section 5.

——Table 1 and Table 2 about here——

The joint probabilities at four possible realizations of the data are summarized for the 12 different models in Table 2. These results show how the joint probabilities depend on the copula families chosen. For example, consider an application with high marginal probabilities for $Y_j = 0$ and a relatively high joint probability that all Y_j are equal to zero even when dependence is taken into account. In this case, a model using all Gumbel pair copulas would be more consistent with the high joint probability that all $Y_j = 0$ than a model with only MTCJ/Clayton pair copulas. The results in Table 2 demonstrate the flexibility of D-vine PCCs but also beg the question of how pair copula building blocks should be selected. We turn our attention to this issue in Section 5.

At this juncture we make some remarks about an alternative, and perhaps at first more intuitive approach, proposed by Smith and Khaled (2010). Here, a different extension of the D-vine is made, with the discrete response determined by a threshold model with an underlying random vector that follows a continuous D-vine. We emphasize, that this approach yields a different, albeit numerically close, model to our own, primarily since conditioning on $\mathbf{V}_{\setminus h} = \mathbf{v}_{\setminus h}$ have different meanings in the continuous and discrete case. Although Smith and Khaled (2010) make valuable contributions in developing several Markov chain Monte Carlo (MCMC) schemes to estimate this model, we believe the construction of an efficient sampling scheme with data augmentation remains a highly challenging problem. In any case, our model retains significant computational advantages over any alternative that requires estimation by MCMC with data augmentation.

4 Estimation and Simulation Study

4.1 Estimation

We consider two maximum likelihood based alternatives to estimating our discrete D-vine model. The first is based on a 2-step inference functions for margins (IFM) approach (for a full summary of both 2-step and multi-step approaches for both continuous and discrete data, see Joe (1997) and references therein). Let the model for the j^{th} margin imply a marginal distribution function $F_j(y_{ij}; \boldsymbol{\mu}_j)$, where $\boldsymbol{\mu}_j$ are the marginal parameters and the subscript i denotes that we observe a sample $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})'$ for $i = 1, 2, \ldots, n$. In the first step, maximum likelihood estimates of the marginal parameters, denoted by $\hat{\boldsymbol{\mu}}_j$, are estimated for all j one at a time, ignoring dependence with the other margins. We then define $u_{ij}^{+} := F_j(y_{ij}|\hat{\boldsymbol{\mu}}_j)$ and $u_{ij}^{-} := F_j(y_{ij} - 1|\hat{\boldsymbol{\mu}}_j)$. In the second step, the copula parameters are estimated by maximum likelihood. For the 2-step estimator the likelihood is computed by using u_{ij}^{+} and u_{ij}^{-} in place of F_j^{+} and F_j^{-} in the algorithm outlined in Appendix A and taking the product over n observations.

Alternately, both the copula and marginal parameters can be estimated jointly by maximum likelihood. As this potentially requires the numerical optimization of a high-dimensional likelihood function, good starting values are required. Starting values for the marginal parameters are obtained by following the first step of the IFM approach. Starting values for the copula parameters can be found by computing empirical Kendall's τ on either u_{ij}^{+} or on estimates of $F_{Y_j|\mathbf{V}}(Y_j \leq y_j|\mathbf{V} = \mathbf{v})$ which act as 'pseudo' data. These can then be transformed back to the copula parameter using a known bijection.[2] While similar methods have been used for parameter estimation in the continuous case by Aas et al. (2009), Czado et al. (2010) and Haff et al. (2010), we recommend using this method only to obtain starting values, as empirical estimates of Kendall's τ are not robust in the discrete case (see Denuit and Lambert (2005)).

As the algorithm to compute the probability mass function of a discrete D-vine PCC is rather involved, confidence intervals cannot be obtained by analytical evaluation of the Hessian. Furthermore, for continuous vine PCCs, numerical evaluation of the Hessian has been shown to give negative variance estimates in some cases (see Aas et al. (2009), Min and Czado (2010)). Consequently, we propose nonparametric bootstrapping via re-sampling with replacement as the preferred method for computing confidence intervals (for an introduction to the bootstrap see Efron and Tibshirani (1993)). Throughout this paper, the number of bootstrapped samples is set equal to the sample size of the data. Although bootstrapping can be computationally intensive, we remind the reader that a significant advantage of our model is fast computation of the likelihood. Although, we recognize that credible intervals may be faster to compute using Bayesian estimation, as was shown in the continuous case

[2]for example for the bivariate Gaussian copula $\theta = (2/\pi)\arcsin(\tau)$.

by Min and Czado (2010), we feel MCMC estimators for our model lie beyond the scope of this paper.

4.2 Bernoulli Margins

We now conduct a simulation study to investigate the efficacy of our estimation procedure including the bootstrapped confidence intervals. For this study, we simulate 100 replications of data from each of the 12 models summarized in Table 1 with a sample size of 300. The data are simulated using an algorithm for generation outlined in more detail in Appendix B. We choose Bernoulli distributed margins as this represents the maximum amount of discretization possible. Although we have 15 model parameters, there are only 32 different realizations of $\mathbf{Y}$ that can be observed, and as our sample size is quite small, in most simulated datasets only around 20 unique values of $\mathbf{Y}$ are observed. This simulation study therefore represents a highly challenging environment for estimating the discrete D-vine model. To allow for comparisons between different copulas, we parameterize all models in terms of Kendall's τ. For each parameter and each model, we evaluate the parameter estimates, the relative bias and the root mean square error averaged over the 100 replications. We also compute the proportion of replications where the true parameter value falls inside a 95% bootstrapped confidence interval. We report results for both 2-step and joint estimation to assess whether joint estimation improves the overall quality of estimates.

——Table 3 and Table 4 about here——

Table 3 summarizes results for model 1A, 2B, and 3C allowing for the comparison across different marginal models and the overall level of dependence, while Table 4 compares models 4A, 4B and 4C, allowing for the comparison across different bivariate copula building blocks. From these results we make the following observations. Joint estimates are generally of a higher quality than their 2-step counterparts, but only slightly so. This is an encouraging result as 2-step estimation is simpler and faster, particularly for more complicated marginal models. For model 4C the opposite is observed with joint estimation exhibiting worse bias and RMSE than a 2-step IFM approach. This occurs as the number of unique observations of $\mathbf{y}$ lies between 10 and 15 for some replications, which is less than the overall number of parameters for joint estimation. In this case, the likelihood will be flat in certain regions and estimates become stuck at initial values. We also note that for all models, estimating the copula parameters becomes more difficult for higher order trees. A potential solution to both these problems is to truncate the D-vine, which shall be discussed further in Section 4.3. Overall, however, we conclude that our estimator is reliable over a wide range of modeling scenarios.

4.3 Poisson Margins

To demonstrate the applicability of our model to high-dimensions we conduct one further simulation example. We generate a single dataset of sample size 1000 from a 20-dimensional model with Poisson margins, each with mean 10. To highlight its complexity, we note that this model has 181 parameters.

It is advantageous in such large applications to uncover a more parsimonious dependence structure by identifying whether a subset of pair copula building blocks are equivalent to the independence copula. In the special case where independence pair copulas are used in trees $l+1, \ldots, m-2, m-1$ we say the D-vine is 'truncated after the l^{th} tree' or alternately that the level of truncation of the D-vine is l. Truncating a D-vine after the l^{th} tree is consistent with an l^{th} order Markovian dependence structure. Algorithms for selecting the level of truncation in a D-vine are considered by Brechmann et al. (2010) in the continuous case.

In our 20-dimensional example, we set the pair copulas in the first two trees to Gaussian copulas with Kendall's $\tau = 0.3$ and truncate after the second tree. We estimate the model assuming a level of truncation at 1,2,3 and 4 trees as well as estimating the full model. We compute the AIC for each of these models and select the level of truncation by choosing the model with the lowest AIC. This model selection strategy has been shown to be an effective 'quick and dirty' algorithm in the continuous case by Brechmann et al. (2010). The results of this procedure are summarized in Table 5.

——Table 5 about here——

The AIC improves drastically when going from estimation with truncation after one tree to estimation with truncation after two trees (the true model) and deteriorates slowly thereafter. To test the robustness of this result we simulated 30 replications of the data and truncation after 2 vines was selected every time. The table also demonstrates the advantages in computing time that can be achieved by truncation. We also note that even the full model takes roughly 8 hours to estimate on a 8 AMD Opetron 8384 Quad Core CPU. This is relatively short considering that an alternative based on taking finite differences would require 1000×2^{20} evaluations of the closed form of a multivariate copula to compute the likelihood function just once, while techniques based on computing rectangle probabilities would require the evaluation of a 20-dimensional integral.

Having revealed the potential benefits of model selection for high-dimensional data, we conclude this section with a brief discussion of more sophisticated model selection techniques that have been implemented for vine PCCs in the continuous case. Brechmann et al. (2010) investigate a variety of strategies for the truncation of vines, that incorporate hypothesis tests for non-nested models both for the D-vine and for more general vine structures. Smith et al. (2010) and Min and Czado (2011) develop Bayesian methods that select whether individual pair copulas are independent or otherwise. By applying their methods to real data, they respectively uncover Markovian dependence in intra-day electricity load and in the term structure of interest rates swaps by identifying independent pair copulas in the higher order trees of a D-vine. Although these sophisticated techniques lie beyond the scope of this paper we believe their application to the discrete setting would be an interesting avenue for future research. In particular, it would be interesting to compare sparse vine PCCs, with alternative approaches to modeling high-dimensional discrete data such as random effects models and latent factor models.

5 Applications

In this section we consider two applications of a D-vine PCC to real discrete data. We note that our priority is not to undertake a thorough applied analysis but to demonstrate the advantages and potential of our model, as well as to make a first attempt at addressing some model selection issues for discrete vines. The first application comes from finance microstructure, specifically count data from order books. Here we fit generalized Poisson models with covariates that have an impact both on the mean and dispersion parameter. Our aim here is to show how a copula approach facilitates the straightforward construction of multivariate distributions from sophisticated marginal models. We carry out two-step estimation to show how inferences on the dependence structure can be made using an approach that is fast and simple. The second application is to a survey of headache severity measured at four different times of the day. Since a large number of covariates is available, we consider ordered probit regression on the margins, with copula and marginal parameters estimated jointly. In this application we show how, in addition to making inferences about the dependence structure, joint estimation of the marginal and copulas parameters can also lead to different inferences being made on the significance of some covariates, compared to the case where dependence is ignored.

5.1 Application to Order Books Data

The data for our first application come from an order book of E.ON stocks traded on the Deutsche Börse between the $2^{nd} - 9^{th}$ January, 2004. The data are counts of market orders and limit orders placed on the buy side within consecutive 60 second intervals. To remove the distorting effects of auctions, two six minute intervals from the beginning of the day and from lunchtime, as well as missing data, are excluded from the sample, giving 2489 observations overall. Market orders are an order to purchase a given quantity of stock at the best current price and in our dataset they are subdivided depending on their size relative to the volume available at the best price on the sell (ask) side of the market. Large market orders (LMO) have volume that exceeds the volume at the best ask price, intermediate market orders (IMO) have exactly the same volume as at the best ask and small orders (SMO) have less volume than is available at the best ask. Limit orders, which represent an order to buy stock only once the price reaches a certain level are also subdivided into three categories. A limit order is in the money (INLO) if in falls inside the bid ask spread, a limit order is at the money

(ATLO) if it is made at the prevailing best bid price, while behind the money limit orders (BEHLO) are made below the prevailing best bid price. For a more specific analysis of order book data we recommend Biais et al. (1995), and Grammig et al. (2005).

For the multivariate modeling of order book data we propose (SMO-IMO-LMO-INLO-ATLO-BEHLO) as an ordering for a 6-dimensional D-vine. Large market orders and limit orders in the money are adjacent to one another in the D-vine as they are both aggressive market actions and we therefore expect a high dependence between them. The ordering in the D-vine of market (and limit) orders with respect to one another is intuitive.

We carry out 2-step estimation based on the IFM approach, with Generalized Poisson regression models (see Consul (1989) and Czado et al. (2007)) fitted to the margins in the first step. The regressors include the volume at the best ask, volume at the best bid, the bid-ask spread, the volatility of returns (defined as the root mean squared 1-minute log returns taken over the previous 5 minutes) and the momentum of returns (defined is a log return over the previous 5 minutes). Log link and shifted log link functions were used for the mean and dispersion parameter respectively. These variables are lagged to ensure they are predetermined, thus avoiding endogeneity issues. Lags of the dependent variable (denoted Lagged Dep. in Table 6) were also used, yielding an observation-driven time-series model. In addition to this model, we also estimated marginal models that include a zero-inflation parameter that depends on covariates along the lines of Czado et al. (2007). However, as the market for E.ON stocks is quite liquid, we found little evidence for zero inflation, and we restrict our attention to the case with no zero-inflation parameter. The estimates for the marginal parameters are summarized in Table 6.

——Table 6 about here——

We now turn our attention to the second step of estimation in the IFM method, namely, estimating the copula parameters. We consider using the Gaussian, MTCJ/Clayton and Gumbel copulas as pair copula building blocks in the first tree of our vine, while all other copulas were assumed to be Gaussian. These 3 copulas were chosen because they give different shapes; they are consistent with no tail dependence, lower tail dependence and upper tail dependence respectively. Also, for these copulas, there is a simple one-to-one correspondence between the usual parametrization and Kendall's τ allowing for comparison between pair copulas in vines that use a combination of different copulas as building blocks. By only selecting copulas on the first tree, we need only consider $3^5 = 243$ different models compared to 3^{15} models had we not restricted all pair copulas from the second tree onwards to be Gaussian. In any case, we expect to find the strongest dependence in the first tree. We allow for negative dependence in the MTCJ/Clayton and Gumbel by rotating them by 90 degrees when dependence is negative[3]. We choose the model with the highest maximised value of the likelihood, hereafter referred to as the 'best' model. Since all models have the same number of parameters, this corresponds to choosing the model according to popular information criteria such as the AIC and BIC. While we recognise the potential of more sophisticated methods, this approach represents a simple but practical first attempt at model selection.

The 'best' model, with a log-likelihood of -27103, has Gumbel-Gumbel-Gaussian-Gumbel-Gumbel copulas on the first tree. To further investigate the improvement in the goodness of fit of our selected model we conduct tests given by Vuong (1989) and Clarke (2007) for non-nested models. Both test the null hypothesis that the two models are equally valid against the alternative that the model with the higher log likelihood is to be preferred over the other. We make two model comparisons. First, we compare the 'best' model to the model with the worst fit according to the information criteria, which has a first tree with MTCJ/Clayton-MTCJ/Clayton-Gumbel-MTCJ/Clayton-MTCJ/Clayton copulas. This model will in the sequel be called the 'worst' model and has a log-likelihood of -27418. Here, both the Vuong and Clarke tests give p-values < 0.0001 thus rejecting the null that both models are equally valid. Second, we compare the 'best' model to a model constructed using only Gaussian pair copulas which, with a maximised log-likelihood of -27107, is the 5^{th} best model overall. We note here that this model is numerically close, but not exactly equivalent to a model based on a multivariate Gaussian copula for discrete margins. When comparing the 'best' model to the vine comprised of only Gaussian pair copula building blocks, the Vuong and Clarke tests yield p-values of 0.38 and < 0.0001 respectively. Overall, these comparisons provide evidence

[3] we refer to these constructions as "Rotated Gumbel" and "Rotated MTCJ/Clayton" copulas

that the correct choice of pair copulas can lead to models that are significantly better in a statistical sense.

The estimates for the copula parameters for the 'best' model as well as 95% bootstrapped confidence intervals are summarized in Table 7. To compare results across different classes of copula, estimates of Kendall's τ are also provided in parentheses. We observe that within each tree, the highest dependence is found in pairs that contain in the money limit orders (INLO), while pairs with INLO in the conditioning set are either statistically or economically insignificant. This seems to imply that INLO is in some sense the key variable driving the dependence structure between our count series. In light of this result we believe it may also be worthwhile in a future project to apply different vine structures, such as a C-vine with INLO as the first node, to this dataset.

——Table 7 about here——

5.2 Longitudinal Data on Headache intensity

The data from our second application were obtained from a survey of sufferers of chronic headaches. The full pooled dataset includes 5609 daily observations from 135 respondents who keep a log of the severity of their headaches, allocating a score ranging from 0 for no headache, to 5 for an intense headache. The number of days reported by each respondent varies from 7 to 381 days. Data were recorded at four different times of the day, motivating a 4-dimensional D-vine PCC model, with an intuitive ordering of the margins (morning-afternoon-evening-night) that allows intra-day dependence in headache severity to be investigated. Covariates that measure both individual attributes of the patients and important meteorological variables were also available, motivating an ordered probit regression for marginal models. The covariates selected after exploratory data analysis are summarized in Table 8. More details on this dataset can be found in Varin and Czado (2010).

——Table 8 about here——

As we are mainly interested in intra-day dependence rather than dependence between different days, we only use observations from Mondays and Thursdays. In doing so we make the reasonable assumption that dependence in headache severity is negligible over periods longer than at least three days. After removing all observations where at least some entries are missing we obtain a reduced sample of 621 observations. We also assume that our covariates measuring individual attributes capture all individual specific effects. While, this assumption may be violated, we reiterate that the application here is primarily demonstrative. The development of a multivariate random effects model with a time series component along the lines of Varin and Czado (2010), motivated by this specific application would be an interesting future project, but lies beyond the scope of this paper.

We follow a similar procedure for model selection as in Section 5.1, however, as there are only 6 pair copulas in a 4-dimensional model it is feasible to enumerate all $3^6 = 729$ possible models. The highest value of the log-likelihood is -2608 and is achieved by a combination of pair copulas summarized in Table 9. Once again we refer to this as the 'best' model and use the Vuong and Clarke test to make comparisons with the vine that gives the lowest value of the log likelihood (a value of -2977) and to a vine constructed using only Gaussian copulas (the 8^{th} best model, giving a log-likelihood of -2620). Similar results are obtained as in Section 5.1, the p-values for the Vuong test are < 0.0001 and 0.32 for the comparison of the 'best' model to the 'worst' model and all Gaussian vine respectively, while for the Clarke test both p-values are < 0.0001.

——Table 9 about here——

Also reported in Table 9 are the estimates of the copula parameters with the bounds of a 95% bootstrapped confidence interval (the same results reparametrized in terms of in Kendall's τ are given in parentheses). As expected, the dependence is strongest in the first tree of the D-vine. Interestingly enough, although the dependence in the last tree is the weakest, the copula parameter $\theta_{MN|AE}$ is significantly different from zero and there is no apparent advantage to truncating the D-vine for this model.

Finally Table 10 summarizes the marginal parameter estimates for the ordered probit regression. The statistically significant coefficients at a 95% level of significance are highlighted in bold. We note that these estimates are based on joint estimates and therefore take intra-day dependence into account. In general, a major advantage of taking dependence into account in regression modeling is that more efficient estimates are obtained, which in turn can lead to more powerful significance tests on the coefficients. For comparison we also estimated the marginal models assuming intra-day independence. Joint maximum likelihood perturbs the coefficient estimates and in particular, the coefficients of β_5 in the morning and β_{15} in the night were significant for the joint estimation but not significant when independence is assumed. Overall, we feel this demonstrates that our model, in addition to uncovering interesting dependence properties, can also have an impact on the inferences made on marginal coefficients in regression modeling.

——Table 10 about here——

6 Conclusions

Our broad aim in this paper was to develop a modeling framework for multivariate discrete data that is both flexible and has significant computational advantages. We chose a copula approach due to its generality; marginal models can be selected with great freedom. In this paper alone we have used probit, ordered probit, Poisson and generalized Poisson distributions for the margins. As a framework for constructing multivariate copula models, we developed a discrete analogue to vine PCCs. We showed how the choice of different parametric copulas as building blocks has an impact on joint probabilities thus demonstrating the flexibility of a vine copula approach. We develop fast algorithms for generating from and computing the probability mass function of a D-vine PCC with discrete margins. We showed that our model can be easily estimated without having to take 2^m finite differences of multivariate copula functions, compute high-dimensional integrals, or resort to less efficient composite likelihood methods. We test the quality of both 2-step and joint maximum likelihood estimation in a simulated setting. The advantages, disadvantages and kind of inferences that can be made for both estimation methods were investigated through application to two interesting datasets from financial market micro-structure and medical statistics.

Although we feel that our contribution is a major step forward in the modeling of multivariate discrete data, we discuss three important open questions in more detail. First, as stated in Section 3 an alternative vine PCCs for discrete data has been introduced by Smith and Khaled (2010) based on a latent vector that is distributed as a continuous D-vine. It would be interesting to compare the joint probabilities from this model with those from our own model, for the same choices of parameter values and pair copulas. Consequently, we could determine whether the model introduced in this paper dominates the alternative with respect to its flexibility as well as with regard to computational considerations.

Second, methods for model selection must be developed to truly take advantage of the flexibility and potential for sparsity inherent in vine PCCs. Recognizing the importance of this issue we suggested using information criteria both to reduce the dimensionality of large models by truncating vines and to select between the different families used as pair copula building blocks. While we believe that our use of information criteria constitutes an important first step in addressing the issue of model selection, we acknowledge that more sophisticated approaches should be attempted. These could be based either on fast, simple diagnostics, or in a Bayesian framework could be based on reversible jump MCMC.

Third, as was the case for vine PCCs with continuous margins, we believe that there may be significant advantages to estimating our model in a Bayesian framework. Furthermore, the modular nature of the Gibbs sampler may facilitate the development of more advanced multivariate models. In particular we believe that joint estimation of marginal and copula parameters for random effects models and for parameter driven time series model could be easier to develop in a Bayesian context. As the likelihood for our PCC is fast to compute, we do not need to resort to data augmentation, and as such we believe that MCMC based estimation of our model may have higher computational and sampling efficiency than is usually the case with Bayesian estimators for discrete data copula models.

Overall, our contribution is to propose an approach for modeling multivariate discrete data that is general, flexible, fast and can be extended to higher dimensions. This paper also poses several new research questions but simultaneously provides direction towards possible solutions to these problems. For these reasons, we believe that vine copula approaches have great potential as an area for future research and represent the way forward in the modeling of discrete multivariate data.

Acknowledgments

Anastasios Panagiotelis is a Alexander von Humboldt Post-Doctoral Fellow while the work of Claudia Czado is partially suppoted by German Research Foundation Grant (CZ86_1_3). Harry Joe is supported by an NSERC Discovery grant. The numerical computations were performed on a Linux cluster supported by DFG grant INST 95/919-1 FUGG.

References

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

Bedford, T. and Cooke, R. (2001). Probability density decomposition for conditionally dependent random variables, *Ann. Math. Artf. Intell.* **32**: 245–268.

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

Biais, B., Hillion, P. and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* **52**: 655–681.

Brechmann, E., Czado, C. and Aas, K. (2010). Truncated and simplified regular in high dimensions with application to financial data. Submitted for Publication.

Cameron, A., Li, T., Trivedi, P. and Zimmer, D. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts, *Econometrics Journal* **7**: 566–584.

Clarke, K. (2007). A simple distribution free test for non-nested model selection, *Polit. Anal.* **13**: 347–363.

Consul, P. (1989). *Generalized Poisson Distributions*, Statistics Textbooks and Monographs, Marcel Dekker Inc., New York.

Czado, C. (2010). Pair-copula constructions of multivariate copulas, *in* P. Jaworki, F. Durante, W. Härdle and W.Rychlik (eds), *Workshop on Copula Theory and its Applications*, Springer, pp. 93–109.

Czado, C., Erhardt, V., Min, A. and Wagner, S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates., *Statistical Modelling* **7**: 125–153.

Czado, C., Schepsmaier, U. and Min, A. (2010). Maximum likelihood estimation of mixed C-vines with application to exchange rates. Submitted for Publication.

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

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data, *Journal of Multivariate Analysis* **93**: 40–57.

Dissmann, J. (2010). Statistical inference for regular vines and application. Diploma Thesis, Chair of Mathematical Statistics, Technische Universität München.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Embrechts, P., Lindskog, F. and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management, *in* S. Rachev (ed.), *Handbook of Heavy Tailed Distributions in Finance*, Elsevier.

Fischer, M., Köck, C., Schlüter, S. and Weigert, F. (2009). An empirical analysis of multivariate copula models, *Quantitative Finance* **9**: 839–854.

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

Genz, A. (1992). Numerical computation of multivariate normal probabilities, *Journal of Computational and Graphical Statistics* **1**: 141–149.

Grammig, J., Heinen, A. and Rengifo, E. (2005). Trading activity and liquidity supply in a pure limit order book market. EFA 2005 Moscow Meetings Paper.

Haff, H., Aas, K. and Frigessi, A. (2010). On the simplified pair-copula construction - simply useful or too simplistic?, *Journal of Multivariate Analysis* **101**: 1296–1310.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer Verlag, New York.

Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations, *Journal of the American Statistical Association* **90**: 957–964.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters, *in* L. Rüschendorf, B. Schweizer and M. Taylor (eds), *Distributions with Fixed Marginals and Related Topics*, Institute of Mathematical Statistics, Hayward, California.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

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

Johnson, N., Kotz, S. and Balakrishnan, N. (1997). *Multivariate Discrete Distributions*, Wiley, New York.

Kurowicka, D. and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, Wiley Series in Probability and Statistics, Wiley, Chichester.

Kurowicka, D. and Joe, H. (eds) (2011). *Dependence Modeling: Vine Copula Handbook*, World Scientific Publishing Company, London.

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

Min, A. and Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions, *Journal of Financial Econometrics* **8**: 511–546.

Min, A. and Czado, C. (2011). Bayesian model selection for d-vine pair-copula constructions, *Canadian Journal of Statistics* **To appear**.

Nelsen, R. (2006). *An Introduction to Copulas (2nd. Ed.)*, Springer Series in Statistics, Springer, New York.

Nikoloulopoulos, A.K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data, *Statistics in Medicine* **27**: 6393–6406.

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

Patton, A. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* **2**: 130–168.

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

Sklar, A. (1959). Fonctions de répartition à n dimensions et leur marges, *Publ. Inst. Statist. Univ. Paris* **8**: 229–231.

Smith, M., Gan, Q. and Kohn, R. (2011). Modeling dependence using skew t copulas: Bayesian inference and applications, *Journal of the Applied Econometrics* **To Appear**.

Smith, M. and Khaled, M. (2010). Estimation of copula models with discrete margins. Working Paper.

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

Varin, C. and Czado, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data, *Biostatistics* **11**: 127–138.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statistica Sinica* **21**: 5–42.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* **57**: 307–333.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis, *Canadian Journal of Statistics* **33**: 335–356.

Appendix A

The following algorithm can be used to compute the likelihood of a D-vine. To assist the reader, this algorithm is also presented pictorially in Figure 2 . To reiterate, we let, $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_m)'$ be a $m \times 1$ discrete valued random vector. For simplicity and without loss of generality we assume $\mathbf{Y} \in \mathbb{N}^m$. Let $\mathbf{i}$ be a vector of integers such that $\mathbf{Y}_i$ is a subset of $\mathbf{Y}$ corresponding to margins given by the elements of $\mathbf{i}$, for example if $\mathbf{i} = \{2, 3, 4\}$ then $\mathbf{Y}_i := \{Y_2, Y_3, Y_4\}$. We use the following notation to denote some important probabilities.

$$
\begin{aligned}
F^+_{g|\mathbf{i}} &:= \Pr\left(Y_g \le y_g | \mathbf{Y}_i = \mathbf{y}_i\right), \; F^-_{g|\mathbf{i}} := \Pr\left(Y_g \le y_g - 1 | \mathbf{Y}_i = \mathbf{y}_i\right) \\
f_{g|\mathbf{i}} &:= \Pr\left(Y_g = y_g | \mathbf{Y}_i = \mathbf{y}_i\right)
\end{aligned}
$$

Note that if $\mathbf{i}$ is empty then we simply compute marginal probabilities. We use the following notation for the copula functions evaluated at four different values.

$$
\begin{aligned}
C^{++}_{gj|\mathbf{i}} &:= C_{gj|\mathbf{i}}\left(F^+_{g|\mathbf{i}}, F^+_{j|\mathbf{i}}\right), \quad C^{+-}_{gj|\mathbf{i}} := C_{gj|\mathbf{i}}\left(F^+_{g|\mathbf{i}}, F^-_{j|\mathbf{i}}\right) \\
C^{-+}_{gj|\mathbf{i}} &:= C_{gj|\mathbf{i}}\left(F^-_{g|\mathbf{i}}, F^+_{j|\mathbf{i}}\right), \quad C^{--}_{gj|\mathbf{i}} := C_{gj|\mathbf{i}}\left(F^-_{g|\mathbf{i}}, F^-_{j|\mathbf{i}}\right)
\end{aligned}
$$

Our algorithm to compute the probability mass function corresponding to a discrete D-vine is as follows

1. For $d = 1, \ldots, m$, evaluate F^+_d, F^-_d and $f_d = F^+_d - F^-_d$.

2. For $d = 1, \ldots, m-1$ evaluate $C^{++}_{d,d+1} = C_{d,d+1}\left(F^+_d, F^+_{d+1}\right)$, $C^{+-}_{d,d+1} = C_{d,d+1}\left(F^+_d, F^-_{d+1}\right)$, $C^{-+}_{d,d+1} = C_{d,d+1}\left(F^-_d, F^+_{d+1}\right)$ and $C^{--}_{d,d+1} = C_{d,d+1}\left(F^-_d, F^-_{d+1}\right)$.

3. for $d = 1, m-2$

   (a) Evaluate $F^+_{d|d+1} = \frac{C^{++}_{d,d+1} - C^{+-}_{d,d+1}}{f_{d+1}}$, $F^-_{d|d+1} = \frac{C^{-+}_{d,d+1} - C^{--}_{d,d+1}}{f_{d+1}}$, $f_{d|d+1} = F^+_{d|d+1} - F^-_{d|d+1}$. Note these will be used as the first arguments when computing the copula functions $C_{d,d+2|d+1}$ in step (c) below.

   (b) Evaluate $F^+_{d+2|d+1} = \frac{C^{++}_{d+1,d+2} - C^{-+}_{d+1,d+2}}{f_{d+1}}$, $F^-_{d+2|d+1} = \frac{C^{+-}_{d+1,d+2} - C^{--}_{d+1,d+2}}{f_{d+1}}$ and $f_{d+2|d+1} = F^+_{d+2|d+1} - F^-_{d+2|d+1}$. Note these will be used as the second arguments when computing the copula functions $C_{d,d+2|d+1}$ in step (c) below.

(c) Evaluate $C^{ab}_{d,d+2|d+1}\left(F^{a}_{d|d+1}, F^{b}_{d+2|d+1}\right)$ for $a, b \in +, -$

4. for $t = 3, \ldots, m-1$ and $d = 1, \ldots, m-t$

   (a) Evaluate first argument of copula functions $C_{d,d+t|d+1,\ldots,d+t-1}$ to be computed in step (c) below:

      (i) Evaluate $F^{+}_{d|d+1,\ldots,d+t-1} = \frac{C^{++}_{d,d+t-1|d+1,\ldots,d+t-2} - C^{+-}_{d,d+t-1|d+1,\ldots,d+t-2}}{f_{d+t-1|d+1,\ldots,d+t-2}}$

      (ii) Evaluate $F^{-}_{d|d+1,\ldots,d+t-1} = \frac{C^{-+}_{d,d+t-1|d+1,\ldots,d+t-2} - C^{--}_{d,d+t-1|d+1,\ldots,d+t-2}}{f_{d+t-1|d+1,\ldots,d+t-2}}$

      (iii) Evaluate $f_{d|d+1,\ldots,d+t-1} = F^{+}_{d|d+1,\ldots,d+t-1} - F^{-}_{d|d+1,\ldots,d+t-1}$

   (b) Evaluate second argument of copula functions $C_{d,d+t|d+1,\ldots,d+t-1}$ to be computed in step (c) below:

      (i) Evaluate $F^{+}_{d+t|d+1,\ldots,d+t-1} = \frac{C^{++}_{d+1,d+t|d+2,\ldots,d+t-1} - C^{-+}_{d+1,d+t|d+2,\ldots,d+t-1}}{f_{d+1|d+2,\ldots,d+t-1}}$

      (ii) Evaluate $F^{-}_{d+t|d+1,\ldots,d+t-1} = \frac{C^{+-}_{d+1,d+t|d+2,\ldots,d+t-1} - C^{--}_{d+1,d+t|d+2,\ldots,d+t-1}}{f_{d+1|d+2,\ldots,d+t-1}}$

      (iii) Evaluate $f_{d+t|d+1,\ldots,d+t-1} = F^{+}_{d+t|d+1,\ldots,d+t-1} - F^{-}_{d+t|d+1,\ldots,d+t-1}$

   (c) Evaluate $C^{ab}_{d,d+t|d+1,\ldots,d+t-1}\left(F^{a}_{d|d+1,\ldots,d+t-1}, F^{b}_{d+t|d+1,\ldots,d+t-1}\right)$ for $a, b \in +, -$ Evaluate copula functions

5. Evaluate $F^{+}_{1|2,\ldots,m} = \frac{C^{++}_{1,m|2,\ldots,m-1} - C^{+-}_{1,m|2,\ldots,m-1}}{f_{m|2,\ldots,m-1}}$, $F^{-}_{1|2,\ldots,m} = \frac{C^{-+}_{1,m|2,\ldots,m-1} - C^{--}_{1,m|2,\ldots,m-1}}{f_{m|2,\ldots,m-1}}$ and $f_{1|2,\ldots,m} = F^{+}_{1|2,\ldots,m} - F^{-}_{1|2,\ldots,m}$.

6. The probability mass function is given by $f_{1|2,\ldots,m} \prod_{i=1}^{m-1} f_{m-i+1|2,\ldots,m-i}$.

——Figure 2 about here——

Appendix B

The following algorithm can be used to generate from the D-vine with discrete margins outlined in Section 3. We continue to assume $\mathbf{Y} \in \mathbb{N}$ for clarity of exposition although generating from any discrete domain is possible.

1. Generate $u_1, u_2, \ldots, u_m \sim$ iid Uniform$(0, 1)$

2. Initialize $k_1, k_2, \ldots, k_m$ each to 1

3. Set $G_C \leftarrow \Pr(Y_{m-1} \leq k_{m-1})$.

4. **While** $u_{m-1} > G_C$

   (a) $k_{m-1} \leftarrow k_{m-1} + 1$

   (b) $G_C \leftarrow \Pr(Y_{m-1} \leq k_{m-1})$

5. Set $G_C \leftarrow \Pr(Y_{m-2} \leq k_{m-2} | Y_{m-1} = k_{m-1})$.

6. **While** $u_{m-2} > G_C$

   (a) Set $k_{m-2} \leftarrow k_{m-2} + 1$

   (b) $G_C \leftarrow \Pr(Y_{m-2} \leq k_{m-2} | Y_{m-1} = k_{m-1})$

7. **For** $j = m-3, m-4, \ldots, 1$:

   (a) Set $G_C \leftarrow \Pr(Y_j \leq k_j | Y_{j+1} = k_{j+1}, \ldots, Y_{m-1} = k_{m-1})$.

   (b) **While** $u_j > G_C$

      (i) Set $k_j \leftarrow k_j + 1$

      (ii) Set $G_C \leftarrow \Pr(Y_j \leq k_j | Y_{j+1} = k_{j+1}, \ldots, Y_{m-1} = k_{m-1})$

8. Set $G_C \leftarrow \Pr(Y_m \leq k_m | Y_1 = k_1, \ldots, Y_{m-1} = k_{m-1})$.

9. **While** $u_m > G_C$

   (a) Set $k_m \leftarrow k_m + 1$

   (b) Set $G_C \leftarrow \Pr(Y_m \leq k_m | Y_1 = k_1, \ldots, Y_{m-1} = k_{m-1})$

The probabilities at steps 4, 6, 7(b) and 9 can be computed using some of the steps outlined in Appendix A.

Case	Marginal Probabilities	Dependence	Model	Pair Copulas
1	**Low:** $\Pr(Y_j = 0) = 0.3$	**Low:** Kendall's τ=0.3,0.2,0.1,0.05	1A	All Gaussian
	for $j = 1, \ldots, 5$	for pair copulas corresponding	1B	All MTCJ/Clayton
		to trees $T_1, \ldots, T_4$ respectively	1C	All Gumbel
2	**High:** $\Pr(Y_j = 0) = 0.7$	**Low:** Kendall's τ=0.3,0.2,0.1,0.05	2A	All Gaussian
	for $j = 1, \ldots, 5$	for pair copulas corresponding	2B	All MTCJ/Clayton
		to trees $T_1, \ldots, T_4$ respectively	2C	All Gumbel
3	**Low:** $\Pr(Y_j = 0) = 0.3$	**High:** Kendall's $\tau = 0.7, 0.4, 0.3, 0.2$	3A	All Gaussian
	for $j = 1, \ldots, 5$	for pair copulas corresponding	3B	All MTCJ/Clayton
		to trees $T_1, \ldots, T_4$ respectively	3C	All Gumbel
4	**High:** $\Pr(Y_j = 0) = 0.7$	**High:** Kendall's $\tau = 0.7, 0.4, 0.3, 0.2$	4A	All Gaussian
	for $j = 1, \ldots, 5$	for pair copulas corresponding	4B	All MTCJ/Clayton
		to trees $T_1, \ldots, T_4$ respectively	4C	All Gumbel

Table 1: Summary of parameter values for the 5-dimensional D-vine with Bernoulli margins discussed in Section 3.3 and used for the simulation study in Section 4.2.

$\mathbf{Y}$	All Gaussian	All MTCJ	All Gumbel	Independent
	Case 1: Low marginal prob., Low dependence			
00000	0.0377	0.0482	0.0319	0.0024
01010	0.0097	0.0102	0.0097	0.0132
10101	0.0157	0.0128	0.0173	0.0309
11111	0.3185	0.3835	0.2920	0.1681
	Case 2: High marginal prob, Low dependence			
00000	0.3185	0.2672	0.3603	0.1681
01010	0.0157	0.0190	0.0144	0.0309
10101	0.0097	0.0099	0.0105	0.0132
11111	0.0377	0.0261	0.0437	0.0024
	Case 3: Low marginal prob., High Dependence			
00000	0.1648	0.1839	0.1683	0.0024
01010	0.0028	0.0026	0.0028	0.0132
10101	0.0037	0.0013	0.0047	0.0309
11111	0.5366	0.6267	0.5028	0.1681
	Case 4: High marginal prob., High dependence			
00000	0.5366	0.4553	0.5899	0.1681
01010	0.0037	0.0054	0.0032	0.0309
10101	0.0028	0.0023	0.0040	0.0132
11111	0.1648	0.1610	0.1706	0.0024

Table 2: Some joint probabilities for a 5 dimensional D-vine with Bernoulli margins. The lefthand column describes the realisation of $\mathbf{Y}$, for example '01010' denotes $Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1$ and $Y_5 = 0$. Here, 'Low marginal prob.' refers to cases where $\Pr(Y_j = 0) = 0.3$ for all j, 'High marginal prob.' refers to cases where $\Pr(Y_j = 0) = 0.7$ for all j, 'Low dependence', refers to to cases where Kendall's $\tau = 0.3, 0.2, 0.1, 0.05$ for pair copulas corresponding to the first to fourth trees respectively and 'High dependence' refers to the case where Kendall's $\tau = 0.7, 0.4, 0.3, 0.2$ for pair copulas corresponding to the first to fourth trees respectively. In the right most column the joint probabilities for independent margins are given for comparison.

Parameter	True	2-step Mean	2-step Rel.Bias	2-step RMSE	2-step Cov.	Joint Mean	Joint Rel.Bias	Joint RMSE	Joint Cov.
Model 1A									
τ_{12}	0.3	0.2901	-0.0330	0.0618	0.91	0.2971	-0.0096	0.0609	0.95
τ_{23}	0.3	0.2926	-0.0245	0.0614	0.90	0.2996	-0.0013	0.0597	0.94
τ_{34}	0.3	0.2850	-0.0500	0.0601	0.93	0.2921	-0.0262	0.0586	0.97
τ_{45}	0.3	0.2943	-0.0188	0.0629	0.95	0.3015	0.0050	0.0622	0.97
$\tau_{13\|2}$	0.2	0.1982	-0.0088	0.0575	0.94	0.2004	0.0019	0.0597	0.95
$\tau_{24\|3}$	0.2	0.1956	-0.0220	0.0657	0.91	0.1991	-0.0044	0.0670	0.94
$\tau_{35\|4}$	0.2	0.2001	0.0005	0.0715	0.86	0.2002	0.0009	0.0738	0.88
$\tau_{14\|23}$	0.1	0.1106	0.1059	0.0661	0.93	0.1008	0.0083	0.0663	0.96
$\tau_{25\|34}$	0.1	0.1003	0.0030	0.0615	0.96	0.0894	-0.1058	0.0634	0.96
$\tau_{15\|234}$	0.05	0.0632	0.2635	0.0601	0.95	0.0460	-0.0793	0.0663	0.96
$\Pr(Y_1 = 0)$	0.3	0.2945	-0.0184	0.0263	0.96	0.2944	-0.0188	0.0263	0.96
$\Pr(Y_2 = 0)$	0.3	0.2988	-0.0039	0.0237	0.99	0.2988	-0.0041	0.0237	0.99
$\Pr(Y_3 = 0)$	0.3	0.3022	0.0073	0.0252	0.96	0.3021	0.0070	0.0252	0.97
$\Pr(Y_4 = 0)$	0.3	0.2994	-0.0019	0.0304	0.91	0.2994	-0.0019	0.0304	0.91
$\Pr(Y_5 = 0)$	0.3	0.3039	0.0129	0.0307	0.92	0.3038	0.0125	0.0307	0.91
Model 2A									
τ_{12}	0.3	0.2873	-0.0425	0.0609	0.94	0.2933	-0.0223	0.0593	0.96
τ_{23}	0.3	0.2912	-0.0295	0.0594	0.96	0.2972	-0.0094	0.0580	0.97
τ_{34}	0.3	0.2877	-0.0411	0.0604	0.95	0.2937	-0.0210	0.0595	0.97
τ_{45}	0.3	0.2924	-0.0252	0.0661	0.90	0.2988	-0.0039	0.0659	0.91
$\tau_{13\|2}$	0.2	0.1941	-0.0295	0.0747	0.89	0.1921	-0.0396	0.0740	0.91
$\tau_{24\|3}$	0.2	0.1959	-0.0204	0.0666	0.92	0.1953	-0.0237	0.0605	0.95
$\tau_{35\|4}$	0.2	0.2054	0.0270	0.0647	0.94	0.2070	0.0349	0.0631	0.95
$\tau_{14\|23}$	0.1	0.1036	0.0362	0.0696	0.90	0.0884	-0.1164	0.0711	0.93
$\tau_{25\|34}$	0.1	0.1196	0.1965	0.0718	0.95	0.1068	0.0676	0.0618	0.98
$\tau_{15\|234}$	0.05	0.0661	0.3228	0.0806	0.90	0.0465	-0.0708	0.0734	0.96
$\Pr(Y_1 = 0)$	0.7	0.7005	0.0007	0.0289	0.90	0.7003	0.0004	0.0290	0.90
$\Pr(Y_2 = 0)$	0.7	0.7004	0.0006	0.0261	0.93	0.7002	0.0003	0.0262	0.94
$\Pr(Y_3 = 0)$	0.7	0.7044	0.0062	0.0257	0.94	0.7042	0.0060	0.0257	0.95
$\Pr(Y_4 = 0)$	0.7	0.7017	0.0024	0.0285	0.92	0.7015	0.0022	0.0284	0.92
$\Pr(Y_5 = 0)$	0.7	0.7084	0.0120	0.0303	0.90	0.7081	0.0116	0.0303	0.91
Model 3A									
τ_{12}	0.7	0.6986	-0.0020	0.0427	0.92	0.6989	-0.0015	0.0419	0.92
τ_{23}	0.7	0.6993	-0.0011	0.0453	0.92	0.6997	-0.0005	0.0443	0.88
τ_{34}	0.7	0.6982	-0.0025	0.0439	0.93	0.6986	-0.0020	0.0430	0.91
τ_{45}	0.7	0.6990	-0.0014	0.0453	0.93	0.6996	-0.0005	0.0443	0.94
$\tau_{13\|2}$	0.4	0.4290	0.0725	0.1048	0.89	0.4304	0.0761	0.1139	0.88
$\tau_{24\|3}$	0.4	0.4179	0.0447	0.0970	0.91	0.4241	0.0602	0.0979	0.93
$\tau_{35\|4}$	0.4	0.4234	0.0586	0.0991	0.92	0.4247	0.0617	0.1121	0.92
$\tau_{14\|23}$	0.3	0.3329	0.1096	0.1137	0.94	0.3384	0.1279	0.1190	0.97
$\tau_{25\|34}$	0.3	0.3138	0.0460	0.1280	0.90	0.3196	0.0652	0.1347	0.91
$\tau_{15\|234}$	0.2	0.2287	0.1436	0.1538	0.91	0.2301	0.1505	0.1619	0.90
$\Pr(Y_1 = 0)$	0.3	0.2971	-0.0098	0.0296	0.90	0.2972	-0.0094	0.0291	0.87
$\Pr(Y_2 = 0)$	0.3	0.2979	-0.0071	0.0276	0.92	0.2981	-0.0064	0.0272	0.91
$\Pr(Y_3 = 0)$	0.3	0.2972	-0.0092	0.0285	0.92	0.2974	-0.0086	0.0284	0.92
$\Pr(Y_4 = 0)$	0.3	0.2994	-0.0019	0.0304	0.91	0.2993	-0.0023	0.0301	0.89
$\Pr(Y_5 = 0)$	0.3	0.2988	-0.0041	0.0295	0.90	0.2990	-0.0034	0.0290	0.89

Table 3: Mean, relative bias (Rel. Bias), root mean square error(RMSE) and coverage (Cov.) of both 2-step and joint estimates for models 1A, 2A and 3A described in Table 1. Results are averaged over 100 replications of data each having sample size 300. Coverage refers to the proportion of simulations where a 95% bootstrapped confidence interval contains the true parameter value.

Parameter	True	2-step Mean	2-step Rel.Bias	2-step RMSE	2-step Cov.	Joint Mean	Joint Rel.Bias	Joint RMSE	Joint Cov.
Model 4A (All Gaussian)									
τ_{12}	0.7	0.7011	0.0016	0.0462	0.88	0.7040	0.0057	0.0456	0.89
τ_{23}	0.7	0.6959	-0.0059	0.0452	0.93	0.6988	-0.0017	0.0439	0.94
τ_{34}	0.7	0.7000	-0.0000	0.0469	0.89	0.7028	0.0040	0.0459	0.93
τ_{45}	0.7	0.6999	-0.0001	0.0523	0.88	0.7028	0.0039	0.0506	0.89
$\tau_{13\|2}$	0.4	0.4264	0.0659	0.0858	0.89	0.4091	0.0226	0.0793	0.91
$\tau_{24\|3}$	0.4	0.4158	0.0394	0.0914	0.88	0.4010	0.0025	0.0853	0.92
$\tau_{35\|4}$	0.4	0.4140	0.0351	0.0808	0.94	0.3963	-0.0093	0.0753	0.95
$\tau_{14\|23}$	0.3	0.3354	0.1181	0.1102	0.88	0.3036	0.0120	0.1022	0.91
$\tau_{25\|34}$	0.3	0.3272	0.0908	0.1004	0.98	0.2950	-0.0165	0.0786	0.98
$\tau_{15\|234}$	0.2	0.2476	0.2380	0.1436	0.91	0.2022	0.0111	0.1053	0.91
$\Pr(Y_1 = 0)$	0.7	0.6998	-0.0003	0.0263	0.95	0.6993	-0.0010	0.0265	0.95
$\Pr(Y_2 = 0)$	0.7	0.7023	0.0032	0.0256	0.96	0.7020	0.0029	0.0252	0.96
$\Pr(Y_3 = 0)$	0.7	0.7042	0.0060	0.0244	0.94	0.7038	0.0055	0.0247	0.94
$\Pr(Y_4 = 0)$	0.7	0.7017	0.0024	0.0285	0.92	0.7014	0.0020	0.0289	0.90
$\Pr(Y_5 = 0)$	0.7	0.7070	0.0100	0.0264	0.97	0.7064	0.0092	0.0263	0.96
Model 4B (All MTCJ/Clayton)									
τ_{12}	0.7	0.7063	0.0091	0.0464	0.92	0.7062	0.0089	0.0463	0.91
τ_{23}	0.7	0.7024	0.0034	0.0428	0.91	0.7026	0.0037	0.0428	0.92
τ_{34}	0.7	0.7060	0.0085	0.0436	0.90	0.7060	0.0086	0.0436	0.91
τ_{45}	0.7	0.7076	0.0109	0.0490	0.88	0.7074	0.0105	0.0488	0.88
$\tau_{13\|2}$	0.4	0.4081	0.0203	0.0831	0.92	0.4080	0.0199	0.0829	0.91
$\tau_{24\|3}$	0.4	0.3963	-0.0091	0.0900	0.93	0.3963	-0.0092	0.0898	0.92
$\tau_{35\|4}$	0.4	0.4047	0.0116	0.0868	0.93	0.4045	0.0112	0.0867	0.93
$\tau_{14\|23}$	0.3	0.3062	0.0207	0.0990	0.93	0.3061	0.0202	0.0989	0.93
$\tau_{25\|34}$	0.3	0.3068	0.0226	0.0852	0.99	0.3067	0.0223	0.0851	0.99
$\tau_{15\|234}$	0.2	0.1840	-0.0798	0.1380	0.94	0.1841	-0.0796	0.1381	0.94
$\Pr(Y_1 = 0)$	0.7	0.7000	0.0000	0.0275	0.91	0.6994	-0.0009	0.0275	0.89
$\Pr(Y_2 = 0)$	0.7	0.7025	0.0036	0.0254	0.93	0.7022	0.0031	0.0252	0.96
$\Pr(Y_3 = 0)$	0.7	0.7036	0.0052	0.0260	0.93	0.7033	0.0047	0.0262	0.93
$\Pr(Y_4 = 0)$	0.7	0.7017	0.0024	0.0285	0.92	0.7012	0.0017	0.0284	0.91
$\Pr(Y_5 = 0)$	0.7	0.7075	0.0108	0.0272	0.92	0.7068	0.0097	0.0270	0.92
Model 4C (All Gumbel)									
τ_{12}	0.7	0.7079	0.0113	0.0473	0.92	0.7100	0.0143	0.0498	0.91
τ_{23}	0.7	0.6973	-0.0038	0.0447	0.95	0.7000	-0.0001	0.0466	0.94
τ_{34}	0.7	0.7019	0.0027	0.0503	0.92	0.7047	0.0068	0.0520	0.91
τ_{45}	0.7	0.7064	0.0091	0.0486	0.90	0.7090	0.0129	0.0506	0.90
$\tau_{13\|2}$	0.4	0.4091	0.0228	0.0908	0.93	0.4213	0.0532	0.1006	0.93
$\tau_{24\|3}$	0.4	0.4139	0.0348	0.0752	0.97	0.4235	0.0588	0.0892	0.96
$\tau_{35\|4}$	0.4	0.3894	-0.0266	0.0865	0.92	0.3990	-0.0024	0.0925	0.93
$\tau_{14\|23}$	0.3	0.2997	-0.0011	0.0958	0.95	0.3131	0.0435	0.1116	0.95
$\tau_{25\|34}$	0.3	0.2999	-0.0002	0.0884	0.96	0.3120	0.0400	0.1076	0.95
$\tau_{15\|234}$	0.2	0.1975	-0.0126	0.0949	0.96	0.2177	0.0886	0.1192	0.96
$\Pr(Y_1 = 0)$	0.7	0.7003	0.0004	0.0263	0.96	0.6999	-0.0001	0.0268	0.96
$\Pr(Y_2 = 0)$	0.7	0.6999	-0.0001	0.0253	0.96	0.7002	0.0003	0.0247	0.95
$\Pr(Y_3 = 0)$	0.7	0.7030	0.0043	0.0238	0.96	0.7025	0.0035	0.0241	0.96
$\Pr(Y_4 = 0)$	0.7	0.7017	0.0024	0.0285	0.92	0.7017	0.0024	0.0287	0.89
$\Pr(Y_5 = 0)$	0.7	0.7071	0.0101	0.0257	0.97	0.7067	0.0096	0.0252	0.99

Table 4: Mean, relative bias (Rel. Bias), root mean square error (RMSE) and coverage (Cov.) of both 2-step and joint estimates for models 4A, 4B and 4C described in Table 1. Results are averaged over 100 replications of data each having sample size 300. Coverage refers to the proportion of simulations where a 95% bootstrapped confidence interval contains the true parameter value.

Level of Truncation	AIC	Computing time (seconds)
1	98447	2759
2	**94341**	**6975**
3	94356	9450
4	94369	11924
19	94490	32252

Table 5: AIC and computing time for 20-dimensional simulated dataset with Poisson margins with mean 10 discussed in Section 4.3. The pair copulas on the first two trees are Gaussian with Kendall's $\tau = 0.3$, while all other pair copulas are the independence copula. The sample size is 1000 and similar results were obtained for 30 replications of data generated from this model.

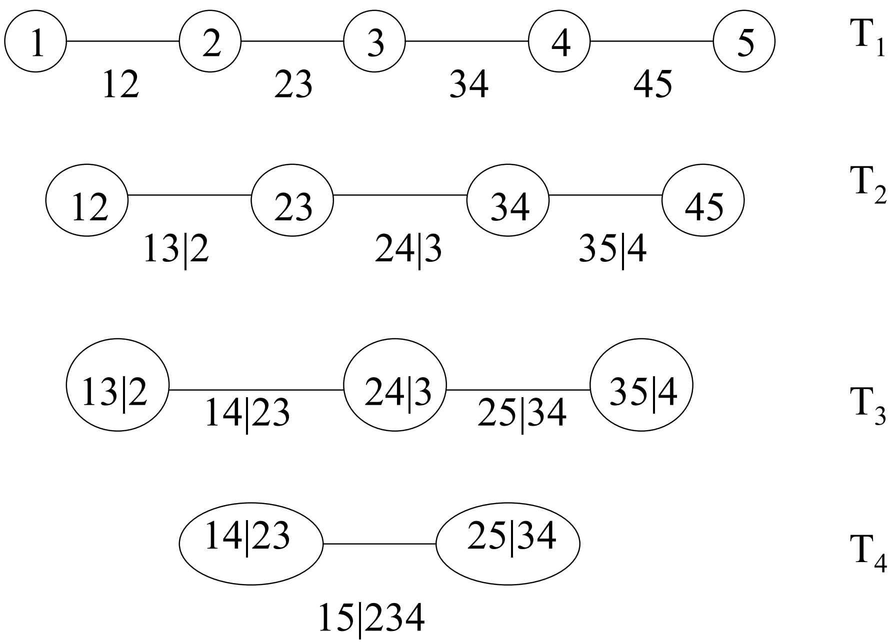

Figure 1: A 5 dimensional D-vine, with T_j denoting the j^{th} tree for $j = 1, 2, 3, 4$

Mean Regression						
Covariate	SMO	IMO	LMO	INLO	ATLO	BEHLO
Intercept	**0.5928**	**-0.7793**	**-0.3750**	**0.7621**	**1.3058**	**2.5163**
Spread	**-0.1900**	**-0.1148**	**-0.2312**	**0.0646**	**0.0492**	0.0023
Bid Vol.	**0.0473**	0.0030	**0.0866**	**0.0378**	-0.0075	**-0.0315**
Ask Vol.	**0.1232**	**-0.1386**	**-0.0996**	-0.0149	0.0238	0.0011
Volatility	**0.1801**	**0.1282**	**0.2783**	**0.2138**	**0.0977**	**0.1071**
Momentum	-0.0314	**-0.0856**	-0.0530	-0.0456	**-0.0338**	-0.0107
Lagged Dep.	**0.2359**	**0.1943**	**0.2047**	**0.1989**	**0.2401**	**0.2719**
Over-Dispersion Regression						
Covariate	SMO	IMO	LMO	INLO	ATLO	BEHLO
Intercept	**-0.8798**	**-1.2778**	**-1.3853**	**-0.7416**	**-0.4841**	**0.4271**
Spread	0.0150	0.0320	0.0644	0.0365	**0.0960**	-0.0081
Bid Vol.	-0.0389	-0.1561	-0.0765	**-0.1710**	-0.0820	-0.0170
Ask Vol.	0.0885	-0.0798	-0.0352	0.0261	-0.0163	0.0322
Volatility	**0.1462**	-0.2078	**0.2548**	**0.1196**	0.0181	-0.0403
Momentum	-0.0589	-0.0231	-0.0479	0.0505	-0.0065	0.0063
Lagged Dep.	**0.2336**	-0.0566	**0.2320**	0.0586	**0.1397**	0.0534

Table 6: Summary of estimates of the marginal parameters for the application in Section 5.1. Recall, that SMO,IMO and LMO correspond to small, intermediate and large market orders respectively, while INLO, ATLO and BEHLO respectively correspond to limit orders in, at and behind the money. Regression effects are included for both the mean regression and over-dispersion regression. Estimates in **bold** denote significance at a 95% confidence level.

Parameter	Copula	ML Estimate	Lower 95% CI	Upper 95% CI
$\theta_{SMO,IMO}$	Gumbel	1.162 (0.14)	1.135 (0.12)	1.192 (0.16)
$\theta_{IMO,LMO}$	Gumbel	1.184 (0.16)	1.156 (0.13)	1.217 (0.18)
$\theta_{LMO,INLO}$	Gaussian	0.624 (0.41)	0.600 (0.39)	0.649 (0.43)
$\theta_{INLO,ATLO}$	Gumbel	1.479 (0.32)	1.441 (0.31)	1.517 (0.34)
$\theta_{ATLO,BEHLO}$	Gumbel	1.343 (0.26)	1.311 (0.24)	1.375 (0.27)
$\theta_{SMO,LMO\|IMO}$	Gaussian	0.383 (0.25)	0.348 (0.22)	0.419 (0.27)
$\theta_{IMO,INLO\|LMO}$	Gaussian	0.335 (0.22)	0.295 (0.19)	0.374 (0.24)
$\theta_{LMO,ATLO\|INLO}$	Gaussian	0.074 (0.05)	0.030 (0.02)	0.117 (0.07)
$\theta_{INLO,BEHLO\|ATLO}$	Gaussian	0.400 (0.26)	0.371 (0.24)	0.430 (0.28)
$\theta_{SMO,INLO\|IMO,LMO}$	Gaussian	0.174 (0.11)	0.139 (0.09)	0.210 (0.13)
$\theta_{IMO,ATLO\|LMO,INLO}$	Gaussian	0.142 (0.09)	0.099 (0.06)	0.185 (0.12)
$\theta_{LMO,BEHLO\|INLO,ATLO}$	Gaussian	0.031 (0.02)	-0.007 (-0.00)	0.070 (0.04)
$\theta_{SMO,ATLO\|IMO,LMO,INLO}$	Gaussian	0.161 (0.10)	0.125 (0.08)	0.196 (0.13)
$\theta_{IMO,BEHLO\|LMO,INLO,ATLO}$	Gaussian	0.055 (0.04)	0.013 (0.01)	0.099 (0.06)
$\theta_{SMO,BEHLO\|IMO,LMO,INLO,ATLO}$	Gaussian	0.040 (0.03)	0.002 (0.00)	0.074 (0.05)

Table 7: Summary of estimates for the dependence parameters in the application in Section 5.1 with 95% bootstrapped confidence intervals. Parentheses are for estimated Kendall's τ and corresponding lower/upper confidence intervals.

Covariate		Coding
Education	β_1	High school or less =0/Bachelor or more=1
Employment (Base Case: Full Time)	β_2	Part-time
	β_3	Unemployed/Disability
	β_4	Other
Marital Status	β_5	Single=0/Married=1
Headache Type	β_6	Migraine=0/Mixed migraine and tension=1
Take Analgesics?	β_7	No=0/Yes=1
Take Prophylactics?	β_8	No=0/Yes=1
Years suffered	β_9	Continuous covariate
Years suffered^2	β_{10}	Continuous covariate
Neuroticism Score	β_{11}	Continuous variable
Program Status (Base: Before program)	β_{12}	During program
	β_{13}	After program
Interaction Variable	β_{14}	Neuroticism $\times$ During Program
Interaction Variable	β_{15}	Neuroticism $\times$ After Program
Humidity (Base Case: [0,59.9%])	β_{16}	[60,79.9]
	β_{17}	[80,100]
Change Air Pressure (Base Case: No change)	β_{18}	Low($<$ 1013 kPa) to High($>$ 1013 kPa)
	β_{19}	High($>$ 1013 kPa) to Low($<$ 1013 kPa)

Table 8: Summary of covariates used in application in Section 5.2

Parameter	Pair Copula	ML Estimate	Lower 95% CI	Upper 95% CI
θ_{MA}	Gaussian	0.664 (0.44)	0.549 (0.36)	0.733 (0.49)
θ_{AE}	Gaussian	0.662 (0.46)	0.569 (0.37)	0.760 (0.51)
θ_{EN}	MTCJ/Clayton	2.247 (0.53)	1.546 (0.44)	3.019 (0.60)
$\theta_{ME\|A}$	MTCJ/Clayton	0.423 (0.17)	0.164 (0.08)	0.6530 (0.25)
$\theta_{AN\|E}$	Gaussian	0.439 (0.28)	0.262 (0.17)	0.508 (0.33)
$\theta_{MN\|AE}$	Gaussian	0.244 (0.16)	0.092 (0.06)	0.347 (0.22)

Table 9: Summary of estimates of copula parameters for the headache application in Section 5.2 with 95% bootstrapped confidence intervals also included. Parentheses are for estimated Kendall's τ and corresponding lower/upper confidence intervals. Here, 'M' denotes Morning, 'A' Afternoon, 'E' Evening and 'N' Night, so that $\theta_{MN|AE}$, for example, describes the dependence between headache severity in the morning and night conditional on headache severity in the afternoon and evening.

	MORNING	AFTERNOON	EVENING	NIGHT
β_1	-0.27	0.16	0.31	0.27
β_2	**0.96**	0.31	0.39	0.32
β_3	**1.06**	**1.13**	**1.32**	**1.49**
β_4	0.05	-0.05	0.31	0.01
β_5	-0.35	-0.28	**-0.46**	-0.38
β_6	-0.00	**0.73**	0.24	0.20
β_7	**1.25**	**1.46**	**0.80**	**1.01**
β_8	**0.44**	**0.49**	**0.50**	**0.63**
β_9	0.01	-0.02	0.01	0.00
β_{10}	0.00	0.00	-0.00	-0.00
β_{11}	**0.05**	**0.04**	0.03	0.03
β_{12}	**3.43**	**2.92**	1.69	1.93
β_{13}	**4.37**	2.57	1.10	3.56
β_{14}	**-0.06**	**-0.05**	-0.02	-0.03
β_{15}	**-0.08**	-0.04	-0.02	**-0.06**
β_{16}	-0.12	0.12	0.14	-0.12
β_{17}	-0.17	-0.19	0.17	-0.16
β_{18}	0.38	0.20	0.12	0.22
β_{19}	-0.14	0.01	0.24	0.06

Table 10: Summary of marginal parameter estimates for the D-vine with ordered probit margins in Section 5.2, based on joint estimation with pair copulas chosen as in Table 9. The covariates that correspond to each of the coefficients (β's) are descibed in Table 8. The figures in **bold** have 95% bootstrapped confidence intervals that do not contain 0.

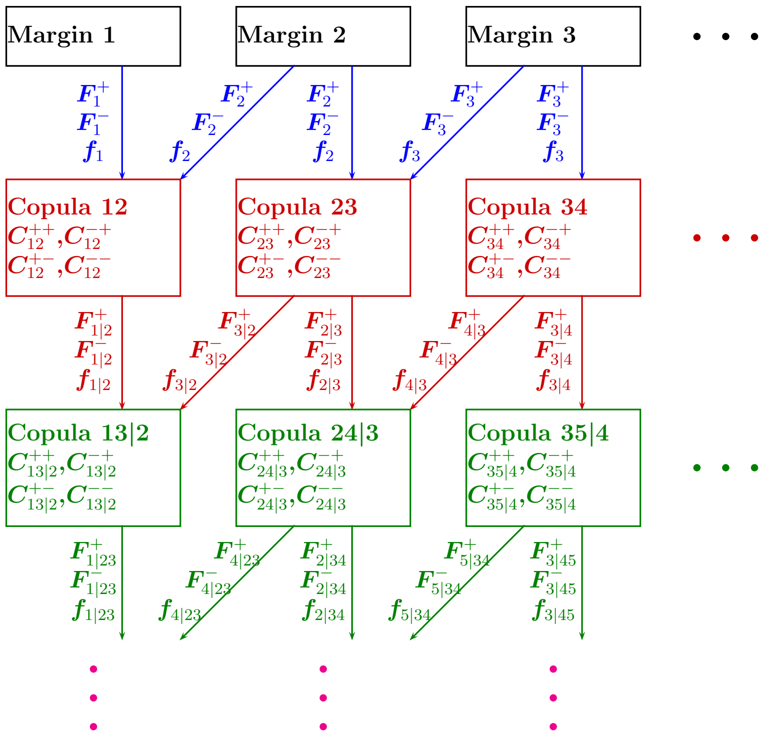

Figure 2: Representation of the algorithm described in Appendix A. Recall $F^{+}_{g|\mathbf{i}} := \Pr(Y_g \leq y_g | \mathbf{Y}_i = \mathbf{y}_i)$, $F^{-}_{g|\mathbf{i}} := \Pr(Y_g \leq y_g - 1 | \mathbf{Y}_i = \mathbf{y}_i)$, $f_{g|\mathbf{i}} := \Pr(Y_g = y_g | \mathbf{Y}_i = \mathbf{y}_i)$ and $C^{ab}_{gj|\mathbf{i}} := C_{gj|\mathbf{i}}\left(F^{a}_{g|\mathbf{i}}, F^{b}_{j|\mathbf{i}}\right)$, for $a = +, -, b = +, -$.