Factor, Structured Factor and Vine Copula Models for Multivariate Social Science Data

Sayed Hasan Kadhem

A thesis submitted for the degree of
Doctor of Philosophy

School of Computer Science
University of East Anglia

June 2022

Factor, Structured Factor and Vine Copula Models for Multivariate Social Science Data

Sayed Hasan Kadhem

June 2022

Abstract

The development of multivariate models with parsimonious dependence is of great interest in a wide range of applications. Two broad frameworks have been considered for parsimonious dependence modelling, namely the latent variable (factor) and copula frameworks. Within these two broad frameworks, we propose several factor models based on copulas for modelling parsimonious dependence structures in multivariate social science data.

We develop factor copula models for mixed continuous and discrete responses where the dependence among the observed variables is explained via a few factors. These are conditional independence models; the observed variables are conditionally independent given the factors.

We also propose the bi-factor and second-order copula models for item response data that can be split into non-overlapping groups, where each group of items has homogeneous dependence. These proposed models fall under the structured factor copula class. Our general models subsume the Gaussian bi-factor and second-order models as special cases and are suitable for capturing different dependencies between and within different groups of observed variables.

Using the vine copula framework, we extend the factor copula models in order to capture any residual dependence. We propose combined factor/truncated vine copula models for item response data. These are conditional dependence models given very few factors. The proposed models can be viewed as a truncated regular vine copula models that involve both observed and latent variables. They allow for flexible construction based on a sequence of bivariate copulas that can provide different tail, asymmetric and non-linear dependence properties.

All the proposed copula models are applied to real datasets and are compared with other relevant benchmark models showing substantial improvement and performance both conceptually and in fit to data.

Statement of publications

I certify that this thesis is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified).

The material presented in Chapter 2 was adapted into a paper published in a peer-reviewed journal:

1. Kadhem, S. H. and Nikoloulopoulos, A. K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, 74(3):365–403.

The material presented in Chapter 3 was adapted into a paper and has been submitted for peer-review. The arXived version of the paper is:

2. Kadhem, S. H. and Nikoloulopoulos, A. K. (2021) Bi-factor and second-order copula models for item response data. *ArXiv e-prints*, arXiv:2102.10660.

R functions for estimation, simulation, model selection and goodness-of-fit of the aforementioned models are part of the R package FactorCopula within the open source statistical environment R:

3. Kadhem, S. H. and Nikoloulopoulos, A. K. (2021). *FactorCopula: Factor, Bi-Factor and Second-Order Copula Models*. R package version 0.8. URL: http://CRAN.R-project.org/package=FactorCopula.

The material presented in Chapter 4 was adapted into a paper and has been submitted for peer- review. The arXived version of the paper is:

4. Kadhem, S. H. and Nikoloulopoulos, A. K. (2022) Factor tree copula models for item response data. *ArXiv e-prints*, arXiv:2201.00339.

To my parents ...

Acknowledgements

I want to express my sincerest appreciation and gratitude to my supervisor, Dr. Aristidis K. Nikoloulopoulos, who has introduced me to copula modelling. I am deeply indebted to him for his sincere help, patience, valuable advice, and unparalleled kindness throughout my Ph.D. journey. Thank you so much for consistently mentoring and guiding me, particularly whenever I felt hopeless with my research! Without your encouragement and help, this research could have never been possible.

I am also extremely grateful to my other supervisor Dr. Beatriz De La Iglesia for her unwavering support, care, and guidance. I have been blessed to have her in my supervisory team, I'm extremely grateful to her.

I would also like to thank Professor Harry Joe (University of British Columbia) for comments that led to an improved presentation of Kadhem and Nikoloulopoulos (2021b) and for the detailed report that I received after the Virtual Vine Copulas seminar organised by the University of British Columbia and Technische Universität München, in which I presented the work in Kadhem and Nikoloulopoulos (2021a).

My sincere appreciation goes to Dr. Katharina Huber, who so generously offered me guidance and advice in so many different ways.

I would also like to thank Dr Irina Irincheeva (University of Bern) and Professor Marc Genton (King Abdullah University of Science and Technology) for providing the Swiss consumption survey dataset.

I would like to acknowledge that the simulations presented in this thesis were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

Finally, I am grateful to my family, especially to my parents for their unconditional love and support throughout the years. Words would never express how grateful I am to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Studying the dependence among multivariate variables is of a great interest in many applications. Association and dependence are interchangeably used in the literature to describe a general relationship of two or more variables. The dependence is usually described via the multivariate normal (MVN) distribution. However, the MVN is not suitable when data display dependence among extreme values and inferences based on multivariate tail probabilities are needed (e.g., Joe et al. 2010). When it is necessary to have copula models with flexible dependence among extreme values, then copulas, is a plausible choice. The theory and application of copulas have become important in finance, insurance and other areas, in order to deal with dependence in the joint tails (e.g., Nikoloulopoulos 2017). Copulas are a useful way to model multivariate data as they account for the dependence structure and provide a flexible representation of the multivariate distribution. Furthermore, as they separate the dependence from the marginal properties, they construct multivariate models with marginal distributions of arbitrary form and allow a wide range of dependence. In fact they allow for flexible modelling of the dependence far from assuming simple linear correlation structures.

An important modelling framework that overcomes the limitations of the MVN is the copula framework. Some desired properties for a parametric family of copulas are (Nikoloulopoulos and Karlis, 2009; Nikoloulopoulos et al., 2012; Nikoloulopoulos, 2013a)

- Wide range of dependence, allowing both perfect positive and negative dependence.
- Flexible dependence, meaning that the number of bivariate marginals is equal to the number of dependence parameters.
- Flexible range of dependence among extreme values.
- Closed form density or cumulative distribution function (cdf) for continuous and discrete data, respectively, and if not of closed-form, then a form that is computationally feasible for estimation.
- Closure property under marginalization, meaning that lower order margins belong to the same parametric family.

Besides the appealing dependence properties of copulas and their popularity in many applications, using simple copulas for multivariate data holds some drawbacks. For example, d-variate Archimedean copulas (McNeil and Nešlehová, 2009) provide only exchangeable dependence with a narrower range of negative dependence for $d > 2$.

To achieve more flexible dependence modelling in high-dimensional data, vine copula models or pair-copula constructions have been proposed (e.g., Joe 1996; Bedford and Cooke 2001, 2002; Kurowicka and Cooke 2006; Kurowicka and Joe 2011; Joe 2014). Vine copulas are a flexible class of models for high-dimensional data

that are constructed from a sequence of bivariate copulas in hierarchies or tree levels. They can accommodate combinations of arbitrary bivariate copulas that have different dependence properties. With appropriate choices of bivariate copulas, vine copulas satisfy all the aforementioned properties except the closure under marginalization property.

Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015) have proposed factor copula models for multivariate continuous and discrete variables, respectively. Factor copulas are vine copulas that involve both observed and latent variables and satisfy all the aforementioned properties including the the closure under marginalization property. In this thesis, we propose several statistical models for dependence modelling using the factor copula framework for multivariate social data. In the forthcoming sections we define (vine) copulas and introduce their basic properties along with the thesis contributions, organisation and structure of the subsequent chapters.

1.1 Basic copula definitions

A copula is a multivariate distribution with standard uniform margins (Nelsen, 2006; Joe, 1997, 2014). In order to provide a precise definition of copulas for the d-variate case, we have first to define the volume of a distribution.

Definition 1.1.1 (Nelsen 2006)**.** *Let $S_1, \ldots, S_d$ be non-empty subsets of $[-\infty, \infty]$, and let H be a d-place real function such that $DomH = S_1 \times \cdots \times S_d$. Let $B = [\mathbf{a}, \mathbf{b}]$ be a d-box all of whose vertices are in $DomH$. Then the H-volume of B is given by*

$$V_H(B) = \sum sgn(\mathbf{c}) H(\mathbf{c}),$$

where the sum is taken over all vertices $\mathbf{c}$ *of* B*, and* $sgn(\mathbf{c})$ *is given by,*

$$sgn(\mathbf{c}) = \begin{cases} 1, & \text{if } c_k = a_k \text{ for an even number of } k\text{'s.} \\ -1, & \text{if } c_k = a_k \text{ for an odd number of } k\text{'s.} \end{cases}$$

Definition 1.1.2 (Nelsen 2006)**.** *A* d*-variate copula is a function* C *from* $[0,1]^d$ *to* $[0,1]$ *with the following properties:*

1. *For every* $\mathbf{u}$ *in* $[0,1]^d$ $C(\mathbf{u}) = 0$ *if at least one coordinate of* $\mathbf{u}$ *is* 0 *and if all coordinates of* $\mathbf{u}$ *are* 1 *except* u_k*, then* $C(\mathbf{u}) = u_k$ *and*

2. *For every* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $[0,1]^m$ *such that* $\mathbf{a} \leq \mathbf{b}$*,* $V_C([a,b]) \geq 0$*.*

The first condition ensures that the marginal distributions are standard uniform. The second condition, often referred to as the rectangular inequality, assures that the copula C is a valid distribution function.

The Sklar's theorem (Sklar, 1959) is central to the theory of copulas, and is the foundation of many, if not most, of the applications of that theory in statistics. Sklar (1959) has elucidated the role that copulas can play in the relationships between multivariate distribution functions and their univariate cdfs.

Theorem 1.1.1 (Sklar 1959)**.** *Let* H *be a* d*-variate cdf with univariate marginal cdfs* $F_1, \ldots, F_d$*. Then there exists a* d*-variate copula* C *such for all* $\mathbf{Y}$

$$H(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)). \tag{1.1}$$

If $F_1, \ldots, F_d$ *are continuous, then* C *is unique, otherwise,* C *is uniquely determined on Range*$F_1 \times \cdots \times$ *Range*F_d*. Conversely, if* C *is a* d*-variate copula and* $F_1, \ldots, F_d$ *are cdfs, then the function* H *defined by (1.1) is a* d*-variate cdf with marginal cdfs* $F_1, \ldots, F_d$*.*

Hence, copulas enable you to break the model building process into two separate steps:

1. Choice of arbitrary marginal distributions:
    - $F_1, \ldots, F_d$ could take different forms;
    - they could involve covariates.
2. Choice of an arbitrary copula function (dependence structure).

The estimation of $F_1, \ldots, F_d$ and C can be done separately.

For every d-variate copula C we know from the Fréchet-Hoeffding inequality (Fréchet, 1951) that copulas are bounded, viz.,

$$\max(u_1 + \cdots + u_d - d + 1, 0) = W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}) = \min(u_1, \ldots, u_d).$$

This can be shown in the bivariate case as follows:

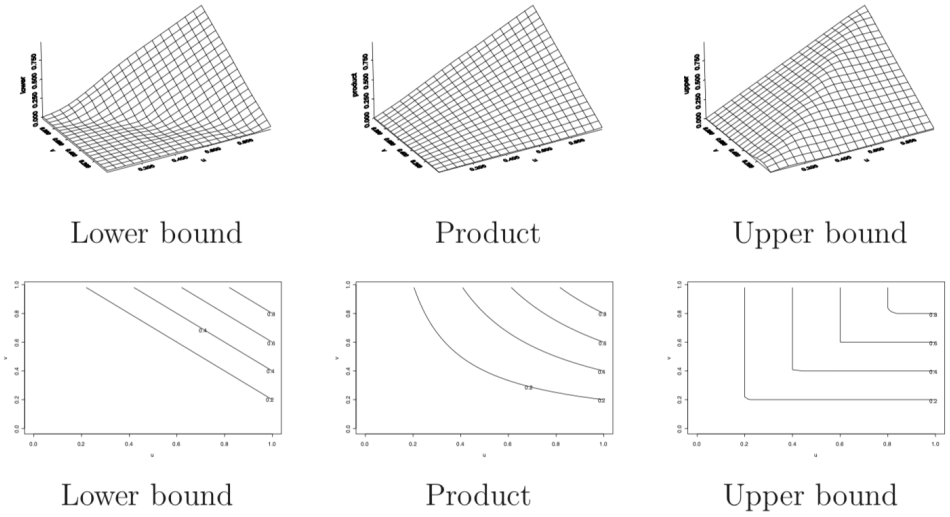
$$\max(u_1 + u_2 - 1, 0) = W(u_1, u_2) \leq C(u_1, u_2) \leq M(u_1, u_2) = \min(u_1, u_2),$$

for $0 \leq u_1, u_2 \leq 1$. While the lower bound W is only a valid copula in the bivariate case, the upper bound M is a valid copula for $d > 2$. We refer to the lower bound and upper bound as countermonotonic and comonotonic copulas, respectively. The comonotonic and countermonotonic copula provides perfect positive and negative dependence, respectively; that is the one variable is strictly increasing and decreasing function of the other, respectively. Another limiting copula is the independence copula, viz.,

$$\Pi(\mathbf{u}) = \prod_{j=1}^{d} u_j,$$

which provides independence between the variables. In Figure 1.1 we depict the independence, comonotonic (upper bound) and countermonotonic (lower bound) copulas.

Figure 1.1: Contour diagrams of the independence (product), comonotonic and countermonotonic copulas.



Since the dependence among random variables is represented by copulas, they provide a natural way to study and measure the association among random variables. Bivariate concordance measures, such as Kendall's τ, Spearman's ρ, and Blomqvist's β (see Chapter 2 of Joe 1997 and Chapter 5 of Nelsen 2006), are copula-based measures of dependence and are margin free, i.e., they do not depend on the univariate margins as the Pearson correlation which is often used in practise as a measure of dependence. For non-normal variates the Pearson correlation can be quite misleading because

- it is only a measure of linear association;
- its value depends on the marginal distributions;
- it can be close to 0 even in case of strong dependence.

Concordance measures of dependence, such as the Kendall's tau, reach

- 1 when $C \equiv M$;
- 0 when $C \equiv \Pi$;
- -1 when $C \equiv W$.

In the forthcoming sections we will define copula-based measures of dependence, such as the Kendall's tau, to describe the dependence in the middle of the data, along with tail dependence and tail order coefficients to describe the dependence in the joint tails of the data.

1.2 Copula-based measures of association

1.2.1 Kendall's τ

Kendall's τ is a common non-parametric measure of concordance, meaning that large (small) values in one variable are associated with large (small) values in another variable, while discordance is when large (small) values of one variable are associated with small (large) values of the other. The Kendall's τ association takes the following form (e.g., Nelsen 2006)

$$\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

The from reveals that Kendall's τ is solely based on the copula C. Bivariate copula parameters have different range, and hence, they are not comparable. In order to make them comparable we use the copula-based Kendall's tau association to quantify the dependence in the middle of the data as they are strictly increasing functions of the the bivariate copula parameters (Nelsen, 2006; Joe, 1997, 2014).

1.2.2 Tail dependence and tail order

Another useful copula-based measure to distinguish among different copula families is tail dependence (Joe, 1993), that is dependence among extreme values. Tail dependence can be used to discriminate different families of bivariate parametric copulas.

A bivariate copula C is *reflection symmetric* if its density satisfies $c(u_1, u_2) = c(1-u_1, 1-u_2)$ for all $0 \leq u_1, u_2 \leq 1$. Otherwise, it is reflection asymmetric often with more probability in the joint upper tail or joint lower tail. *Upper tail dependence* means that $c(1-u, 1-u) = O(u^{-1})$ as $u \to 0$ and *lower tail dependence* means that $c(u, u) = O(u^{-1})$ as $u \to 0$. If $(U_1, U_2) \sim C$ for a bivariate copula C, then $(1-U_1, 1-U_2) \sim \widehat{C}$, where $\widehat{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1-u_1, 1-u_2)$ is the survival or reflected copula of C; this "reflection" of each uniform $U(0,1)$ random variable about $1/2$ changes the direction of tail asymmetry.

Following Hua and Joe (2011), we also define the copula-based tail order coefficients. Under some regularity conditions (e.g., existing finite density in the interior of the unit square, ultimately monotone in the tail), if there exists $\kappa_L(C) > 0$ and some $L(u)$ that is slowly varying at 0^+ $\left(i.e., \frac{L(ut)}{L(u)} \sim 1, \text{ as } u \to 0^+ \text{ for all } t > 0\right)$, then $\kappa_L(C)$ is the *lower tail order* of C. The *upper tail order* $\kappa_U(C)$ can be defined by the reflection of (U_1, U_2), i.e., $\overline{C}(1-u, 1-u) \sim u^{\kappa_U(C)} L^*(u)$ as $u \to 0^+$, where $\overline{C}$ is the survival function of the copula and $L^*(u)$ is a slowly varying function. With $\kappa = \kappa_L$ or κ_U, a bivariate copula has *intermediate tail dependence* if $\kappa \in (1, 2)$, *tail dependence* if $\kappa = 1$, and *tail quadrant independence* if $\kappa = 2$ with $L(u)$ being asymptomatically a constant.

1.3 Parametric families of bivariate copulas

We start by discussing bivariate parametric copulas to allow for a more concrete exposition. Later, we will discuss about d-variate parametric copulas, namely the d-dimensional vine copulas, which are built via successive mixing from $d(d-1)/2$ bivariate linking copulas on trees.

We will consider bivariate parametric copulas that have different tail dependence (Joe, 1993) or tail order (Hua and Joe, 2011). Note that there is a rich literature on bivariate parametric copulas (see e.g., Joe 2014; Nelsen 2006), here we list the most common copula families that capture different dependence structures of multivariate data.

- Reflection symmetric copulas with intermediate tail dependence such as the BVN copula with $\kappa_L = \kappa_U = 2/(1+\theta)$, where θ is the copula (correlation) parameter. The BVN copula cdf is

$$C(u_1, u_2; \theta) = \Phi_2\Big(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\Big), \quad -1 \leq \theta \leq 1,$$

  where Φ is the univariate standard normal cdf and and Φ_2 is the cdf of a BVN distribution with correlation parameter θ.

- Reflection symmetric copulas with tail quadrant independence ($\kappa_L = \kappa_U = 2$), such as the Frank copula with cdf

$$C(u_1, u_2; \theta) = -\theta^{-1} \log\left\{1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right\}, \theta \in (-\infty, \infty)\backslash\{0\}.$$

- Reflection asymmetric copulas with upper tail dependence only such as

  - the Gumbel copula with $\kappa_L = 2^{1/\theta}$ and $\kappa_U = 1$, where θ is the copula parameter.

– the Joe copula with $\kappa_L = 2$ and $\kappa_U = 1$.

The Gumbel and Joe copula cdf is

$$C(u_1, u_2; \theta) = \exp\Big[-\Big\{(-\log u_1)^\theta + (-\log u_2)^\theta\Big\}^{1/\theta}\Big], \quad \theta \geq 1$$

and

$$C(u_1, u_2; \theta) = 1 - \Big\{(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\Big\}^{1/\theta}, \ \theta \geq 1,$$

respectively.

- Reflection symmetric copulas with tail dependence, such as the t_ν copula with $\kappa_L = \kappa_U = 1$. The t_ν copula cdf is

$$C(u_1, u_2; \theta) = \mathcal{T}_2\Big(\mathcal{T}^{-1}(u_1; \nu), \mathcal{T}^{-1}(u_2; \nu); \theta, \nu\Big), \quad -1 \leq \theta \leq 1,$$

where $\mathcal{T}(;\nu)$ is the univariate Student-t cdf with (non-integer) ν degrees of freedom, and $\mathcal{T}_2$ is the cdf of a bivariate Student-t distribution with ν degrees of freedom and correlation parameter θ.

- Reflection asymmetric copulas with upper and lower tail dependence that can range independently from 0 to 1, such as the BB1 and BB7 copulas with $\kappa_L = 1$ and $\kappa_U = 1$. The BB1 and BB7 copula cdf is

$$C(u_1, u_2; \theta, \delta) = \Big[1 + \Big\{(u_1^{-\theta} - 1)^\delta + (u_2^{-\theta} - 1)^\delta\Big\}^{1/\delta}\Big]^{-1/\theta}, \quad \theta > 0, \delta \geq 1$$

and

$$C(u_1, u_2; \theta, \delta) = 1 - \Big[1 - \Big\{(1-\bar{u}_1^\theta)^{-\delta} + (1-\bar{u}_2^\theta)^{-\delta} - 1\Big\}^{-1/\delta}\Big]^{1/\theta}, \ \theta \geq 1, \delta > 0,$$

with $\bar{u}_1 = 1 - u_1$ and $\bar{u}_2 = 1 - u_2$, respectively.

- Reflection asymmetric copulas with tail quadrant independence, such as the BB8 copula with cdf

$$C(u_1, u_2; \theta, \delta) = \delta^{-1}\Big[1 - \Big\{1 - \eta^{-1}[1 - (1 - \delta u_1)^\theta][1 - (1 - \delta u_2)^\theta]\Big\}^{1/\theta}\Big],$$

where $\theta \geq 1, 0 < \delta \leq 1$, and $\eta = 1 - (1 - \delta)^\theta$, or the BB10 copula with cdf

$$C(u_1, u_2; \theta, \delta) = u_1 u_2 \Big\{1 - \delta(1 - u_1^\theta)(1 - u_2^\theta)\Big\}^{-1/\theta}, \quad \theta > 0, 0 \leq \delta \leq 1.$$

The BVN, Frank, and t_ν are comprehensive copulas, i.e., they interpolate between countermonotonicity (perfect negative dependence) to comonotonicity (perfect positive dependence). The other aforementioned parametric families of copulas, namely Gumbel, Joe, BB1, BB7, BB8 and BB10 interpolate between independence and perfect positive dependence. Nevertheless, negative dependence can be obtained from these copulas by considering reflection of one of the uniform random variables on (0, 1). If $(U_1, U_2) \sim C$ for a bivariate copula C with positive dependence, then

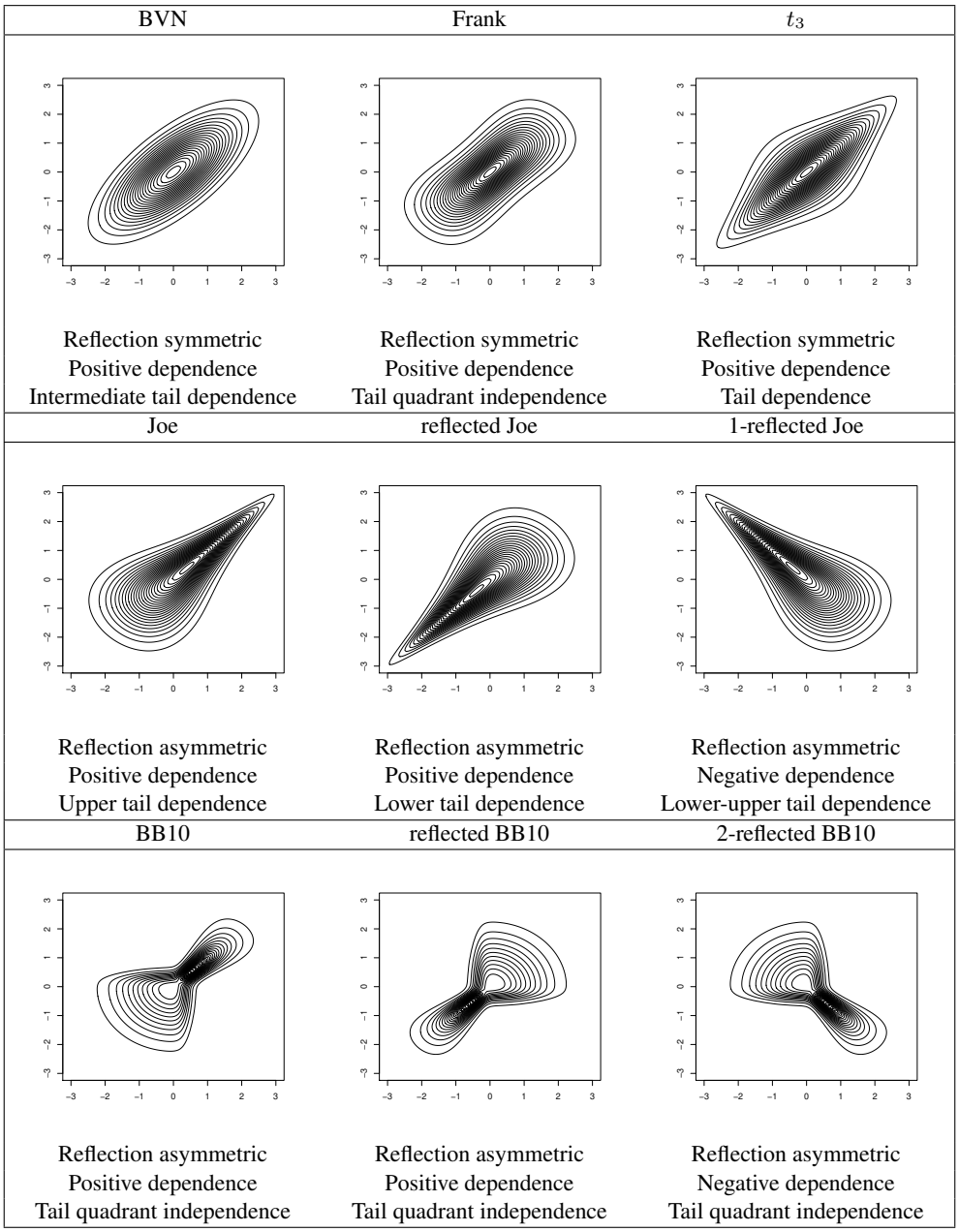
- $(1-U_1, U_2) \sim \widehat{C}^{(1)}$, where $\widehat{C}^{(1)}(u_1, u_2) = u_2 - C(1-u_1, u_2)$ is the 1-reflected copula of C with negative lower-upper tail dependence;
- $(U_1, 1-U_2) \sim \widehat{C}^{(2)}$, where $\widehat{C}^{(2)}(u_1, u_2) = u_1 - C(u_1, 1-u_2)$ is the 2-reflected copula of C with negative upper-lower dependence.

Negative upper-lower tail dependence means that $c(1 - u, u) = O(u^{-1})$ as $u \to 0^+$ and *negative lower-upper tail dependence* means that $c(u, 1 - u) = O(u^{-1})$ as $u \to 0^+$ (Joe, 2011).

In Figure 1.2, to depict the concepts of refection symmetric or asymmetric tail dependence or quadrant tail independence, we show contour plots of the corresponding

copula densities with standard normal margins and dependence parameters corresponding to Kendall's τ value of 0.5 on absolute value. Sharper corners (relative to ellipse) indicate tail dependence.

Figure 1.2: Contour plots of bivariate copulas with standard normal margins and dependence parameters corresponding to Kendall's τ value of 0.5 in absolute value.



1.4 Vine copulas

Vine copula models are flexible tools to analyse dependence structures based on a series of bivariate copulas and have been popular in many application areas (Kurowicka and Joe, 2011).

In order to obtain a valid probability distribution, a $d-$dimensional regular vine has been defined in terms of $d-1$ trees such that $T_1, \ldots, T_{d-1}$ as follows (Bedford and Cooke 2001, 2002):

- T_1 is a tree with nodes $N_1 = 1, \ldots, d$ and edges E_1.
- For $i = 1, \ldots, d-1$, T_i is a tree with nodes $N_i = E_{i-1}$ and edge set E_i. Edges in a tree becomes nodes in the next tree.
- Two edges in tree T_i are joined in tree $T_i + 1$ only if they share a common node in tree T_i. This is known as the proximity condition.

There are two boundary cases of vines, namely, the canonical-vine (C-vine) and drawable-vine (D-vine) models (Nikoloulopoulos et al., 2012). The D-vine model is natural for linear order of time events and longitudinal data (Panagiotelis et al., 2012), while the C-vine model is plausible when the variables there is a (pilot) variable that drives the dependence. In Figure 1.3 we depict the C-vine and D-vine models. In Tree 1 (T_1), the edges correspond to non conditional bivariate copulas, then the edges in T_1 becomes nodes in T_2. These nodes are connected with edges that are given as conditional bivariate copulas, this process continues until the last tree T_4. Note that the arrangement of these variables in this example is arbitrary.

The possibility of parsimonious vine models are obtained via truncation. Truncation of vine models means that the copulas at the higher trees will be set to independence (Brechmann et al., 2012). So the vine models will involve many conditional

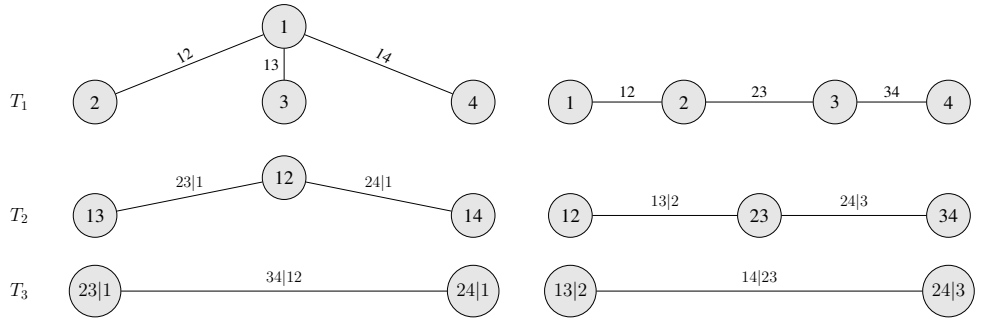


Figure 1.3: Graphical representation of the C-vine (left panel) and D-vine (right panel) copulas with $d = 4$ observed variables.

independent copulas. Truncated vines are often reasonable and sufficient as the dependence amongst the data is mostly explained by the first few trees (Joe et al., 2010).

Factor copula models are truncated C-vine copulas rooted at the latent/unobserved variables. Hence, they are also constructed using a sequence of bivariate copulas that can involve different tail dependence or asymmetry properties. They are more suitable than vine copula models if there exists a latent variable that drives the dependence among the variables.

1.5 Further motivation and thesis contributions

1.5.1 Factor copula models for mixed data

It is very common in the social science (e.g., in surveys) to deal with datasets that have mixed continuous and discrete responses. For example, amount of expenditures and income which are regarded as continuous variables might be included in a survey with other ordinal variables that measure quality of life or depression (Bartholomew et al., 2011). In the literature, two broad frameworks have been considered to model the dependence among such mixed continuous and discrete responses, namely the latent variable and copula frameworks.

There are two approaches for modelling multivariate mixed data with latent variables: the underlying variable approach that treats all variables as continuous by assuming the discrete responses as a manifestation of underlying continuous variables that usually follow the normal distribution (e.g., Muthén 1984; Lee et al. 1992; Quinn 2004); and the response function approach that postulates distributions on the observed variables conditional on the latent variables usually from the exponential family (e.g., Moustaki 1996; Moustaki and Knott 2000; Wedel and Kamakura 2001; Huber et al. 2004; Moustaki and Victoria-Feser 2006). The former method almost invariably assumes that the underlying variables (linked to the observed variables via a threshold process to yield ordinal data and an identity process to yield continuous data) follow a MVN distribution, while the latter assumes that the observed variables are conditionally independent, usually given MVN distributed latent variables. They are equivalent when in the underlying and the response function approach the MVN distribution has a factor and an independence correlation structure, respectively (Takane and de Leeuw, 1987).

The underlying variable approach uses the MVN distribution as a latent model for the discrete responses, and therefore maximum likelihood (ML) estimation requires multidimensional integrations (Nikoloulopoulos, 2013b, 2016); their dimension is equal to the number of observed discrete variables. This is why alternative estimation methods such as the three stage weighted least squares and composite likelihood have been proposed; see e.g., Katsikatsou et al. (2012). The response function approach, with the dependence coming from p latent (unobservable) variables/factors where $p << d$ (the number of observed variables), requires p- rather than d- dimensional integration. Hence, ML estimation is feasible, especially when the number of latent variables is small.

Nevertheless, both approaches are restricted to the MVN assumption for the observed or the latent variables that is not valid if tail asymmetry or tail dependence exists in the mixed data which is a realistic scenario. This occurs when many responses are found in one or both of the extreme ends of the scale and thus the normality assumption is not usually appropriate (Cai et al., 2011). Ma and Genton (2010), Montanari and Viroli (2010), and Irincheeva et al. (2012a) stress that the MVN assumption might not be adequate, and acknowledge that the effect of misspecifying the distribution of the latent variables could lead to biased model estimates and poor fit. To this end, Irincheeva et al. (2012b) proposed a more flexible response function approach by strategically multiplying the MVN density of the latent variables by a polynomial function to achieve departures from normality.

As we have discussed, the underlying variable approach exploits the use of the MVN assumption to model the joint distribution of mixed data. The univariate margins are transformed to normality and then the MVN distribution is fitted to the transformed data. This construction is apparently the MVN copula applied to mixed data (Shen and Weissfeld, 2006; Hoff, 2007; Song et al., 2009; He et al., 2012; Jiryaie et al., 2016), but previous papers (e.g., Quinn 2004) do not refer to copulas as the approach can be explained without copulas.

Smith and Khaled (2012), Stöber et al. (2015) and Zilko and Kurowicka (2016) used vine copulas to model mixed data. Vine copulas have two major advantages over the MVN copula as emphasized in Panagiotelis et al. (2017). The first is that the computational complexity of computing the joint probability distribution function grows only quadratically with d, whereas for the MVN copula the computational complexity grows exponentially with d. The second is that vine copulas are highly flexible through their specification from bivariate parametric copulas with different

tail dependence or asymmetry properties. They have, as special case, the MVN copula, if all the bivariate parametric copulas are bivariate normal (BVN).

In Chapter 2, we extend the factor copula models in Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015) to the case of mixed continuous and discrete responses. Factor copulas are vine copula models that involve both observed and latent variables. Hence, they are highly flexible through their specification from bivariate parametric copulas with different tail dependence or asymmetry properties. The underlying variable approach where the MVN distribution has a p-factor correlation structure or its equivalent, the response function approach where the MVN distribution has an independence correlation structure, is a special case of factor copula models when all the bivariate parametric copulas are BVN (hereafter referred to as the standard factor model).

We tackle issues of particular interest to the social data analyst such as model selection and goodness-of-fit. Model selection in previous papers on factor copula models (Krupskii and Joe, 2013; Nikoloulopoulos and Joe, 2015) was mainly based on simple diagnostics. In addition to simple diagnostics based on semi-correlations (correlations in the lower and upper quadrants of the data), we propose an heuristic method that automatically selects the bivariate parametric copula families. With regard to the issue of goodness-of-fit testing, we propose a technique that is based on the M_2 goodness-of-fit statistic (Maydeu-Olivares and Joe, 2006) in multidimensional contingency tables to overcome the shortage of goodness-of-fit statistics for mixed continuous and discrete response data (e.g., Moustaki and Knott 2000).

1.5.2 Bi-factor and second-order copula models for item response data

Item response data can be defined as the responses of the questions in a survey. They are usually measured in an ordinal scale and constructed to measure unobserved traits

or behavioural characteristics such as extroversion (e.g., Wainer et al. 2007). Datasets with large number of items are often naturally divided into subgroups, in such, each group of items has homogeneous dependence. For example, the well-being (common factor) of patients is usually assessed via items that arise from several sub-domains to assess several group-specific factors such as the depression, anxiety and stress. This special classification of items is also common in educational assessments and referred to as "testlets" (Wainer and Kiely, 1987). It is essential to investigate the structure of the item response data, as implementing factor models on testlet-based items could result in biased estimates and a poor fit (Wang and Wilson, 2005; DeMars, 2006; Zenisky et al., 2002; Sireci et al., 1991; Lee and Frisbie, 1999; Wainer and Thissen, 1996).

To account for the homogeneous dependence in each group of items, Gibbons and Hedeker (1992) and Gibbons et al. (2007) proposed bi-factor models for binary and ordinal response data, respectively. The bi-factor models have become omnipresent in analysing survey items that arise from several sub-domains or groups. They consist of a common factor that is linked to all items, and non-overlapping group-specific factors. The common factor explains dependence between items for all groups, while the group-specific factors explain dependence amongst items within each group. The items are assumed to be independent given the group-specific and common factors.

An alternative way of modelling items that are split into several groups is via the second-order model (e.g., de la Torre and Song 2009; Rijmen 2010), where items are indirectly mapped to an overall (second-order) factor via non-overlapping group-specific (first-order) factors. Second-order models are suitable when the first-order factors are associated with each other, and there is a second-order factor that accounts for the relations among the first-order factors.

The bi-factor and the second-order models are not generally equivalent (Yung et al., 1999; Gustafsson and Balke, 1993; Mulaik and Quartetti, 1997; Rijmen, 2010), unless proportionality constraints are imposed by using the Schmid-Leiman transformation method (Schmid and Leiman, 1957). More importantly, both models are restricted to the MVN assumption for the latent variables, which might not be valid. Nikoloulopoulos and Joe (2015) emphasized that if the ordinal variables in item response data can be thought of as discretizations of latent random variables that are maxima/minima or mixtures of means, then the use of factor models based on the MVN assumption for the latent variables could provide poor fit. More discussion is given in Section 3.1.3.

In the context of item response data, latent maxima, minima and means can arise depending on how a respondent considers specific items. An item might make the respondent think about M past events which, say, have values $W_1, \ldots, W_M$. In answering the item, the subject might take the average, maximum or minimum of $W_1, \ldots, W_M$ and then convert to the ordinal scale depending on the magnitude. The case of a latent maxima/minima can occur if the response is based on a best or worst case. For different dependent items based on latent maxima or minima, multivariate extreme value and copula theory can be used to select suitable distributions for the latent variables. Copulas that arise from extreme value theory have more probability in one joint tail (upper or lower) than expected with a MVN distribution and have latent variables that are maxima/minima instead of means. Even, in the case where the item responses are based on discretizations of latent variables that are means, then it is possible that there can be more probability in both the joint upper and joint lower tail, compared with MVN distributed latent variables. This happens if the respondents consist of a "mixture" population (e.g., different locations or genders). From the theory of elliptical distributions and copulas (McNeil et al. 2005; Joe 2014),

it is known that the multivariate Student-t distribution as a scale mixture of MVN has more dependence in the tails.

Nikoloulopoulos and Joe (2015) have studied factor copula models for item response data and have shown that there is an improvement on the factor models based on the MVN assumption for the latent variables both conceptually and in the fit to the data. This improvement relies on the aforementioned reasons, i.e., items can have more probability in joint upper or lower tail than would be expected with a MVN or items can be considered as discretized maxima/minima or mixtures of discretized means rather than discretized means.

In Chapter 3, we propose copula extensions for bi-factor and second-order models. The construction of the bi-factor copula model exploits the use of bivariate copulas that link the observed variables to the common and group-specific factors. We also propose a heuristic method that automatically selects suitable bivariate copulas, along with goodness-of-fit based on the M_2 statistic (Maydeu-Olivares and Joe, 2006) for the bi-factor and second-order copula models. Note that if there is only one group of items, then the bi-factor model reduces to the 2-factor copula model in Nikoloulopoulos and Joe (2015). Similarly with the bi-factor copula model, we also use bivariate copulas to construct the second-order copula model. In this case, there are bivariate copulas that link the observed to the group-specific factors, and also bivariate copulas that link the group-specific to the second-order factor. To account for the dependence between the observed variables and group-specific factors, each group of variables in fact is modelled using the one-factor copula model proposed by Nikoloulopoulos and Joe (2015). In addition, if there is only one group of items, then the second-order copula model reduces to the one-factor copula model. Hence, the proposed models contain the one- and two-factor copula models in Nikoloulopoulos and Joe (2015) as special cases, while allowing flexible dependence structure for

both within and between group dependence. As a result, the models are suitable for modelling a high-dimensional item response classified into non-overlapping groups.

The proposed models are truncated vine copulas (Brechmann et al., 2012) that involve both observed and latent variables. They provide flexible dependence by selecting arbitrary bivariate linking copulas (Joe et al., 2010) to link the items to latent factors. If the bivariate linking copulas are BVN, then the Gaussian bi-factor and second-order models are special cases of our constructions which are the discrete counterparts of the structured factor copula models introduced by Krupskii and Joe (2015).

1.5.3 Factor tree copula models for item response data

Most factor models are restricted to the conditional independence assumption, where the observed variables are assumed to be conditionally independent given some latent variables. This assumption implies that the dependence amongst the observed variables is fully accounted for by the factors with no remaining dependence. This could lead to biased estimates if the strict assumption of conditional independence is violated (Braeken et al., 2007; Sireci et al., 1991; Chen and Thissen, 1997; Yen, 1993). The conditional independence assumption is violated if there exists local or residual dependence. Mitigating the residual dependence might be achieved by adding more latent variables to the factor model, but at the expense of computational problems and difficulties in interpretation and identification.

To circumvent these problems, the items can be allowed to interrelate by forming a dependence structure with conditional dependence given a few interpretable latent variables. In this way, on the one hand the parsimonious feature of factor models remains intact and any residual dependencies are being taken into account on the other. This can be achieved by incorporating copulas into the conditional distribution of factor models in order to provide a conditional dependence structure given

very few latent variables. Such copula approaches for item response data are proposed by Braeken et al. (2007, 2013) and Braeken (2011) who explored the use of Archimedean copulas or a mixture of the independence and comonotonicity (perfect positive dependence) copulas to capture the residual dependence of traditional item response theory models. Therein simple copulas have been used for subgroups of items that are chosen from the context with homogeneous within-subgroup dependence. This is due to the fact that Archimedean copulas allow only for exchangeable dependence with a narrower range as the dimension increases (McNeil and Nešlehová, 2009).

Without a priori knowledge of obvious subgroups of items that are approximately exchangeable, we will propose a more general residual dependence approach that makes use of truncated regular vine copula models (Brechmann et al., 2012). Within a vine copula specification, no such restrictions need to be made. Regular vine copulas are a flexible class of models that are constructed from a set of bivariate copulas in hierarchies or tree levels (Joe, 1996; Bedford and Cooke, 2001, 2002; Kurowicka and Cooke, 2006; Kurowicka and Joe, 2011; Joe, 2014). A d-dimensional regular vine copula can cover flexible dependence structures, rather than assuming simple linear correlation structures, tail independence and normality (Nikoloulopoulos et al., 2012), through the specification of $d-1$ bivariate parametric copulas at level 1 and $\binom{d-1}{2}$ bivariate conditional parametric copulas at higher levels; at level ℓ for $\ell = 2, \ldots, d-1$, there are $d-\ell$ bivariate conditional copulas that condition on $\ell - 1$ variables. Joe et al. (2010) have shown that in order for a vine copula to have (tail) dependence for all bivariate margins, it is only necessary for the bivariate copulas in level 1 to have (tail) dependence and it is not necessary for the conditional bivariate copulas in levels $2, \ldots, d-1$ to have (tail) dependence. That provides the theoretical justification for the idea to model the dependence in the first level and then just use the independence copulas to model conditional dependence at higher levels without

sacrificing the tail dependence of the vine copula distribution. That is the 1-truncated vine copula has $d-1$ parametric bivariate copulas in the 1st level of the vine and independence copulas in all the remaining levels of the vine (truncated after the 1st level). This truncation, as per the terminology in (Brechmann et al., 2012), provides a parsimonious vine copula model. The 1-truncated vine copula can provide, with appropriately chosen linking copulas, asymmetric dependence structure as well as tail dependence (dependence among extreme values). Joe et al. (2010) have shown that by choosing bivariate linking copulas appropriately, vine copulas can have a flexible range of lower/upper tail dependence and different lower/upper tail dependence parameters for each bivariate margin. Choices of copulas with upper or lower tail dependence are better if the items have more joint upper or lower tail probability than would be expected with the discretized multivariate normal (MVN) model (Muthén, 1978). Note in passing that the discretized MVN distribution is a special case of the vine copula model with discrete margins. If all bivariate copulas are bivariate normal (BVN) in the vine copula model, then the resulting model is the discretized MVN.

To define the conditional independence part of the model we also use truncated vine copulas rather than the traditional factor models for item response in Braeken et al. (2007, 2013) and Braeken (2011). Nikoloulopoulos and Joe (2015) have proposed factor copula models for item response data. These factor models can be explained as truncated canonical vines rooted at the latent variables. The canonical vine is a boundary case of regular vine copulas, which is suitable if there exists a (latent) variable that drives the dependence among the items. For the first factor there are bivariate copulas that couple each item to the first latent variable and for the second factor there are copulas that link each item to the second latent variable conditioned on the first factor (leading to conditional dependence parameters), etc. Factor copula models with appropriately chosen linking copulas will be useful when the items (a) have more probability in joint upper or lower tail than would be expected with a dis-

cretized multivariate normal, or (b) can be considered as discretized maxima/minima or mixtures of discretized means rather than discretized means (Nikoloulopoulos and Joe, 2015).

The proposed parsimonious approach, that requires no priori knowledge of the subgroups of items, can be explained as a truncated regular vine copula model that involves both observed and latent variables; but, more simply, we derive the models as conditional dependence models with a few interpretable latent variables that model the residual dependence of the factor copula model via an 1-truncated vine copula. The factor copula model explains most of the dependence and the remaining dependence can be further accounted for by an 1-truncated vine copula conditioned on the factors. One reason to have residual dependence is when the observed variables do not share homogeneous or common dependence (that arise from the latent variables) with the rest of the observed variables. Alternatively, the bi-factor and second-order models can be used for items that are grouped into non-overlapping groups (e.g., Kadhem and Nikoloulopoulos 2021a; Gibbons et al. 2007; Gibbons and Hedeker 1992). While, the combined factor vine copula models avoid such hurdles when the groups of items are overlapping, not known or difficult to identify. They also avoid violating the conditional independence assumption due to their conditional dependence structure. Brechmann and Joe (2014) and Joe (2018) initiated the study of such conditional dependence models with a unidimensional factor/latent variable for continuous data. The combined 1-factor and 1-truncated vine model for continuous data in Brechmann and Joe (2014) is restricted to Gaussian dependence, but Joe (2018) proposed a combination of an 1-factor copula model with 1-truncated vine copula model with non-Gaussian bivariate copulas. Our models for item response are discrete counterparts of the models in Brechmann and Joe (2014) and Joe (2018) with interpretation and technical details that are quite different and provide an extension to more than one factor.

1.6 Thesis organization

The remainder of the thesis is structured as follows. In Chapter 2, we present the factor copula models for mixed continuous and discrete responses (Kadhem and Nikoloulopoulos, 2021b). Model selection algorithms and goodness-of-fit techniques are also proposed and examined through an extensive simulation study. We also present an application of our methodology to three real datasets.

In Chapter 3 we present copula extensions for bi-factor and second-order models for item response data (Kadhem and Nikoloulopoulos, 2021a) and discuss their relationship with the existing models. Model selection algorithms to select suitable bivariate copulas and goodness-of-fit techniques are proposed. The derivations of the M_2 goodness-of-fit statistic of (Maydeu-Olivares and Joe, 2006) for the bi-factor and second-order copula models are also given. We examine our methodology through an extensive simulation study and also present an application of our methodology to a real dataset.

In Chapter 4, we present combined factor/truncated vine copula models for item response data (Kadhem and Nikoloulopoulos, 2022). These are conditional dependence models with very few interpretable latent variables. In this case, the factor model explains most of the dependence and the remaining dependence is further exploited by an 1-truncated vine copula conditioned on the factors. Model selection algorithms to select suitable vine tree and bivariate copulas are proposed and assessed through an extensive simulation study. We also present an application of our methodology by re-analysing a real dataset.

In Chapter 5, we conclude the thesis with some discussion and future research.

Chapter 2

Factor copula models for mixed data

In this chapter, we present factor copula models for mixed continuous and discrete responses. These are conditional independence models, where observed variables are assumed to be conditionally independent given some latent variables. The construction of the proposed factor copula models exploits the use of bivariate copulas that link the observed to the latent variables. Bivariate copulas other than BVN, with different tail behaviour, can be employed to model tail asymmetry or dependence in the data.

Suitable bivariate parametric copulas are selected using a heuristic method, this is a sequential model selection algorithm that we propose. In order to evaluate the fit of the resulting factor copula models for mixed data, we propose a technique based on the M_2 goodness-of-fit statistic (Maydeu-Olivares and Joe, 2006). The M_2 statistic is based on a quadratic form of the deviations of sample and model-based proportions over all bivariate margins.

We illustrate the proposed methodology by re-analysing three real datasets, and

show that factor copula models with selected copulas (obtained from the model selection algorithm) provide substantial improvements over standard factor models that are based on the normality assumption.

The chapter is organised as follows. Section 2.1 introduces factor copula models for mixed data. Estimation techniques and computational details are provided in Section 2.2. Sections 2.3 and 2.4 propose methods for model selection and goodness-of-fit, respectively. Section 2.5 presents applications of our methodology to three mixed response data sets. Section 2.6 contains an extensive simulation study to gauge the small-sample efficiency of the proposed estimation, investigate the misspecification of the bivariate copulas, and examine the reliability of the model selection and goodness-of-fit techniques. We conclude with a summary in Section 2.8.

2.1 The factor copula model for mixed responses

Although the factor copula models can be explained as truncated canonical vines rooted at the latent variables, we derive the models as conditional independence models, i.e., a response function approach with dependence coming from latent (unobservable) variables/factors. The p-factor model assumes that the mixed continuous and discrete responses $\mathbf{Y} = (Y_1, \ldots, Y_d)$ are conditionally independent given p latent variables $X_1, \ldots, X_p$. In line with Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015), we use a general copula construction, based on a set of bivariate copulas that link observed to latent variables, to specify the factor copula models for mixed continuous and discrete variables. The idea in the derivation of this p-factor model will be shown below for the 1-factor and 2-factor case. It can be extended to $p \geq 3$ factors or latent variables in a similar manner. The evaluation of a p-dimensional integral can be successively performed as we strategically assume that the factors or latent variables are independent.

For the 1-factor model, let X_1 be a latent variable, which we assume to be standard uniform (without loss of generality). From Sklar (1959), there is a bivariate copula $C_{X_1 j}$ such that $\Pr(X_1 \leq x, Y_j \leq y) = C_{X_1 j}\big(x, F_j(y)\big)$ for $0 \leq x \leq 1$ where F_j is the cumulative distribution function (cdf) of Y_j. Then it follows that

$$F_{j|X_1}(y|x) := \Pr(Y_j \leq y | X_1 = x) = \frac{\partial C_{X_1 j}\big(x, F_j(y)\big)}{\partial x}. \tag{2.1}$$

Letting $C_{j|X_1}(F_j(y)|x) = \partial C_{X_1 j}(x, F_j(y))/\partial x$ for shorthand notation and $\mathbf{y} = (y_1, \ldots, y_d)$ be realizations of $\mathbf{Y}$, the density[‡] of the observed data in the 1-factor model case is

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^1 \prod_{j=1}^{d} f_{j|X_1}(y_j|x)\, dx, \tag{2.2}$$

where

$$f_{j|X_1}(y|x) = \begin{cases} C_{j|X_1}\big(F_j(y)|x\big) - C_{j|X_1}\big(F_j(y-1)|x\big) & \text{if} \quad Y_j \quad \text{is discrete;} \\ c_{X_1 j}\big(x, F_j(y)\big) f_j(y) & \text{if} \quad Y_j \quad \text{is continuous,} \end{cases}$$

is the density of $Y_j = y$ conditional on $X_1 = x$; $c_{X_1 j}$ is the bivariate copula density of X_1 and Y_j and f_j is the univariate density of Y_j.

For the 2-factor model, consider two latent variables X_1, X_2 that are, without loss of generality, independent uniform $U(0,1)$ random variables. Let $C_{X_1 j}$ be defined as in the 1-factor model, and let $C_{X_2 j}$ be a bivariate copula such that

$$\Pr(X_2 \leq x_2, Y_j \leq y | X_1 = x_1) = C_{X_2 j}\big(x_2, F_{j|X_1}(y|x_1)\big),$$

[‡]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 integer values this is the counting measure on the set of possible outcomes, for continuous margins we consider the Lebesgue measure in $\mathbb{R}$.

where $F_{j|X_1}$ is given by (2.1). Then for $0 \leq x_1, x_2 \leq 1$,

$$\Pr(Y_j \leq y|X_1 = x_1, X_2 = x_2) = \frac{\partial}{\partial x_2} \Pr(X_2 \leq x_2, Y_j \leq y|X_1 = x_1)$$

$$= \frac{\partial}{\partial x_2} C_{X_2 j}\big(x_2, F_{j|X_1}(y|x_1)\big) = C_{j|X_2}\big(F_{j|X_1}(y|x_1)|x_2\big).$$

The density of the observed data in the 2-factor model case is

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2 j|X_1}\big(x_2, y_j|x_1\big)\, dx_1 dx_2, \tag{2.3}$$

where $f_{X_2 j|X_1}(x_2, y|x_1)$

$$= \begin{cases} C_{j|X_2}\big(F_{j|X_1}(y|x_1)|x_2\big) - C_{j|X_2}\big(F_{j|X_1}(y-1|x_1)|x_2\big) & \text{if } Y_j \text{ is discrete;} \\ c_{jX_2;X_1}\big(F_{j|X_1}(y|x_1), x_2\big) c_{X_1 j}\big(x_1, F_j(y)\big) f_j(y) & \text{if } Y_j \text{ is continuous.} \end{cases}$$

Note that the copula $C_{X_1 j}$ links the jth response to the first latent variable X_1, and the copula $C_{X_2 j}$ links the jth response to the second latent variable X_2 conditional on X_1. In our general statistical model there are no constraints on the choice of the parametric marginal F_j or copula $\{C_{X_1 j}, C_{X_2 j}\}$ distribution.

2.1.1 Semi-correlations to detect tail dependence or tail asymmetry

Choices of copulas with upper or lower tail dependence are better if the observed variables have more probability in joint upper or lower tail than would be expected with the standard factor model. This can be shown with summaries of correlations in the upper joint tail and lower joint tail.

For continuous variables, although copula theory uses transforms to standard uniform margins $U_j = F_j(Y_j)$, we convert to normal scores $Z_j = \Phi^{-1}(U_j)$ to check deviations from the elliptical shape that would be expected with the BVN copula (Nikoloulopoulos et al., 2012). For notational ease, let $C_{2|1} = C_{2|1}\big(0.5|\Phi(z)\big)$, and

$c_{12} = c\big(\Phi(z_1), \Phi(z_2)\big)$, then the correlations of normal scores in the upper and lower tail (hereafter semi-correlations) are defined as (Joe, 2014, page 71):

$$\begin{aligned}
\rho_N^+ &= \text{Cor}\Big(Z_{j_1}, Z_{j_2} | Z_{j_1} > 0, Z_{j_2} > 0\Big) \\
&= \frac{\int_0^\infty \int_0^\infty z_1 z_2 \phi(z_1)\phi(z_2) c_{12} dz_1 dz_2 - \Big(\int_0^\infty z\phi(z)\Big(1 - C_{2|1}\Big)dz\Big)^2 / C(0.5, 0.5)}{\int_0^\infty z^2\phi(z)\Big(1 - C_{2|1}\Big)dz - \Big(\int_0^\infty z\phi(z)\Big(1 - C_{2|1}\Big)dz\Big)^2 / C(0.5, 0.5)}; \\
\rho_N^- &= \text{Cor}\Big(Z_{j_1}, Z_{j_2} | Z_{j_1} < 0, Z_{j_2} < 0\Big) \\
&= \frac{\int_{-\infty}^0 \int_{-\infty}^0 z_1 z_2 \phi(z_1)\phi(z_2) c_{12} dz_1 dz_2 - \Big(\int_{-\infty}^0 z\phi(z) C_{2|1} dz\Big)^2 / C(0.5, 0.5)}{\int_{-\infty}^0 z^2\phi(z) C_{2|1} dz - \Big(\int_{-\infty}^0 z\phi(z) C_{2|1} dz\Big)^2 / C(0.5, 0.5)},
\end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the univariate normal cdf and density, respectively. Note in passing that for the BVN copula $\rho_N^+ = \rho_N^-$ and has a closed form; see (Joe, 2014, page 71).

From the above expressions, it is apparent that the normal scores semi-correlations depend only on the copula C of (U_{j_1}, U_{j_2}). Table 2.1 has semi-correlations for all the aforementioned bivariate parametric copulas with $\tau = \{0.3, 0.5, 0.7\}$. From the table we can see that $\rho_N^+ = \rho_N^-$ for any reflection symmetric copula, while they are different for any reflection asymmetric one. If there is stronger upper (lower) tail dependence than with the BVN, then the upper (lower) semi-correlation is larger.

The population versions ρ_N^+, ρ_N^- also apply when the variables Y_j are ordinal. Under the univariate probit model (Agresti, 2010, Section 3.3.2) Z_j are standard normal underlying latent variables, such that

$$Y_j = y_j \quad \text{if} \quad \alpha_{y_j - 1, j} \leq Z_j \leq \alpha_{y_j j},\ y_j = 1, \ldots, K_j, \tag{2.4}$$

Table 2.1: Lower semi-correlations ρ_N^-, upper semi-correlations ρ_N^+, lower tail dependence λ_L, and upper tail dependence λ_U, with $\tau = \{0.3, 0.5, 0.7\}$ for 1-parameter and 2-parameter bivariate copulas.

Bivariate copula	τ	θ	δ	ρ_N^-	ρ_N^+	λ_L	λ_U
BVN	0.3	0.45		0.23	0.23	0.00	0.00
	0.5	0.71		0.47	0.47	0.00	0.00
	0.7	0.89		0.75	0.75	0.00	0.00
t_3	0.3	0.45		0.45	0.45	0.29	0.29
	0.5	0.71		0.61	0.61	0.45	0.45
	0.7	0.89		0.80	0.80	0.66	0.66
Frank	0.3	2.92		0.15	0.15	0.00	0.00
	0.5	5.74		0.32	0.32	0.00	0.00
	0.7	11.41		0.60	0.60	0.00	0.00
Joe	0.3	1.77		0.05	0.58	0.00	0.52
	0.5	2.86		0.14	0.78	0.00	0.73
	0.7	5.46		0.37	0.92	0.00	0.86
Gumbel	0.3	1.43		0.16	0.46	0.00	0.38
	0.5	2.00		0.36	0.67	0.00	0.59
	0.7	3.33		0.64	0.85	0.00	0.77
BB1	0.3	0.50	1.14	0.43	0.25	0.30	0.17
	0.5	0.35	1.71	0.52	0.59	0.31	0.50
	0.7	1.33	2.00	0.85	0.72	0.77	0.59
BB7	0.3	1.40	0.40	0.28	0.37	0.18	0.36
	0.5	1.50	1.57	0.66	0.42	0.64	0.41
	0.7	4.00	2.00	0.73	0.85	0.71	0.81
BB8	0.3	3.92	0.60	0.10	0.22	0.00	0.00
	0.5	4.51	0.80	0.20	0.52	0.00	0.00
	0.7	6.89	0.90	0.41	0.84	0.00	0.00
BB10	0.3	1.60	0.83	0.18	0.09	0.00	0.00
	0.5	2.50	0.98	0.43	0.19	0.00	0.00
	0.7	10.00	1.00	0.25	0.66	0.00	0.00

where K_j is the number of categories of Y_j and $a_{1j}, \ldots, a_{K_j-1,j}$ are the univariate cutpoints (without loss of generality, we assume $\alpha_{0j} = -\infty$ and $\alpha_{K_j j} = \infty$). Note in passing that for binary variables ($K_j = 2$) the calculation of the semi-correlations is meaningless as the binary variables have no tail asymmetries.

The sample versions of ρ_N^+, ρ_N^- are sample linear (when both variables are continuous), polychoric (when both variables are ordinal), and polyserial (when one vari-

able is continuous and the other is ordinal) correlations in the joint lower and upper quadrants of the two variables. The sample polychoric and polyserial correlation is defined as

$$\hat{\rho}_N = \text{argmax}_\rho \sum_{i=1}^{n} \log\Big(\Phi_2(\alpha_{y_{i1}}, \alpha_{y_{i2}}; \rho) - \Phi_2(\alpha_{y_{i1}-1}, \alpha_{y_{i2}}; \rho) \\ -\Phi_2(\alpha_{y_{i1}}, \alpha_{y_{i2}-1}; \rho) + \Phi_2(\alpha_{y_{i1}-1}, \alpha_{y_{i2}-1}; \rho)\Big),$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the BVN cdf with correlation ρ and

$$\hat{\rho}_N = \text{argmax}_\rho \sum_{i=1}^{n} \log\left\{\phi(z_{i1})\left(\Phi\Big(\frac{\alpha_{y_{i2}} - \rho z_{i1}}{(1-\rho^2)^{1/2}}\Big) - \Phi\Big(\frac{\alpha_{y_{i2}-1} - \rho z_{i1}}{(1-\rho^2)^{1/2}}\Big)\right)\right\}$$

with $z_{ij} = \Phi\Big(\frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(Y_{ij} \leq y_{ij})\Big)$, respectively.

2.2 Estimation

We use a two-stage copula modelling approach toward the estimation of a multivariate model that borrows the strengths of the semi-parametric and inference function for margins (IFM) approach in Genest et al. (1995) and Joe (2005), respectively. Suppose that the data are y_{ij}, $j = 1, \ldots, d$, $i = 1, \ldots, n$, where i is an index for individuals or clusters and j is an index for the within-cluster measurements. For $i = 1, \ldots, n$, we start from a d-variate sample $y_{i1}, \ldots, y_{id}$ from which d estimators $F_1(y_{i1}), \ldots, F_d(y_{id})$ can be obtained. We use these to transform the $y_{i1}, \ldots, y_{id}$ sample into a uniform sample $u_{i1} = F_1(y_{i1}), \ldots, u_{id} = F_d(y_{id})$ on $[0, 1]^d$ and then fit the factor copula model at the second step. For continuous and discrete data y_{ij}, we use non-parametric and parametric univariate distributions, respectively, to transform the data y_{ij} into copula data $u_{ij} = F_j(y_{ij})$, i.e., data on the uniform scale. Hence

our proposed approach, in line with the approaches in Genest et al. (1995) and Joe (2005), can be regarded as a two-step approach on the original data or simply as the standard one-step ML method on the transformed (copula) data.

2.2.1 Univariate modelling

For continuous random variables, we estimate each marginal distribution non-parametrically by the empirical distribution function of Y_j, viz.

$$F_j(y_{ij}) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(Y_{ij} \leq y_{ij}) = R_{ij}/(n+1),$$

where R_{ij} denotes the rank of Y_{ij} as in the semi-parametric estimation of Genest et al. (1995) and Shih and Louis (1995). Hence we allow the distribution of the continuous margins to be quite free and not restricted by parametric families.

Nevertheless, rank-based methods cannot be used for discrete variables with copulas (Genest and Nešlehová, 2007). Hence, for both ordinal and count variables we have chosen realistic parametric models:

- For an ordinal response variable Y_j we use the univariate probit model in (2.4). The ordinal response Y_j is assumed to have density

$$f_j(y_j; \boldsymbol{\gamma}_j) = \Phi(\alpha_{y_j j}) - \Phi(\alpha_{y_j - 1, j}),$$

  where $\boldsymbol{\gamma}_j = (a_{1j}, \ldots, a_{K_j - 1, j})$ is the vector of the univariate cutpoints.

- For a count response variable Y_j we use the negative binomial distribution (Lawless, 1987). It allows for over-dispersion and its probability mass function is

$$f_j(y_j; \boldsymbol{\gamma}_j) = \frac{\Gamma(\xi_j^{-1} + y_j)}{\Gamma(\xi_j^{-1})\ y_j!} \frac{\mu_j^y \xi_j^y}{(1 + \xi_j^{-1})^{\xi_j^{-1} + y_j}}, y_j = 0, 1, 2, \ldots, \mu_j > 0,\ \xi_j > 0,$$

where $\boldsymbol{\gamma}_j = \{\mu_j, \xi_j\}$ is the vector with the mean and dispersion parameters. In the limit $\xi \to 0$ the negative binomial reduces to Poisson, which belongs to the exponential family of distributions and it is the only distribution for count data that existing latent variable models for mixed data can accommodate.

To this end, for a discrete random variable Y_j, we approach estimation by maximizing the univariate log-likelihoods

$$\ell_j(\boldsymbol{\gamma}_j) = \sum_{i=1}^{n} \log f_j(y_{ij}; \boldsymbol{\gamma}_j)$$

over the vector of the univariate parameters $\boldsymbol{\gamma}_j$. That is equivalent with the first step of the IFM method in Joe (1997, 2005). In line with the IFM method, if one uses a misspecified univariate model for the discrete responses at the first step, then the estimation of the copula parameters at the second step deteriorates as demostrated in Kim et al. (2007). Nevertheless, there is no "correct specification" of the margins or copula for data analysis. If one does a proper analysis of the univariate margins for goodness-of-fit, then the proposed two-stage (or IFM) method should be fine. Kim et al. (2007) have "true univariate distributions for simulations" and "specified univariate distributions for estimation" that were very far apart and unrealistic, because the difference of the two is easily detected without too much data.

2.2.2 Copula modelling

After estimating the univariate marginal distributions we proceed to estimation of the dependence parameters. For the 1-factor and 2-factor models, we let $C_{X_1 j}$ and $C_{X_2 j}$ be parametric bivariate copulas, say with dependence parameters θ_j and δ_j, respectively. Let also $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_j, \theta_j : j = 1, \ldots, d\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_j, \theta_j, \delta_j : j = 1, \ldots, d\}$ to denote the set of all parameters for the 1- and 2-factor model, respectively. Estimation

can be achieved by maximizing the joint log-likelihood

$$\ell_{\mathbf{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\mathbf{Y}}(y_{i1}, \ldots, y_{id}; \boldsymbol{\theta}). \tag{2.5}$$

over the copula parameters θ_j or δ_j, $j = 1, \ldots, d$, with the univariate parameters/distributions fixed as estimated at the first step of the proposed two-step estimation approach. The estimated parameters can be obtained by using a quasi-Newton (Nash, 1990) method applied to the logarithm of the joint likelihood. This numerical method requires only the objective function, i.e., the logarithm of the joint likelihood, while the gradients are computed numerically and the Hessian matrix of the second order derivatives is updated in each iteration. The standard errors (SEs) of the estimates can be obtained via the gradients and the Hessian computed numerically during the maximization process. These SEs are adequate to assess the flatness of the log-likelihood. Proper SEs that account for the estimation of univariate parameters can be obtained by maximizing the joint likelihood in (2.5) at one step over $\boldsymbol{\theta}$.

For factor copula models numerical evaluation of the joint density $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ can be easily done using Gauss-Legendre quadrature (Stroud and Secrest, 1966). To compute one-dimensional integrals for the 1-factor model, we use the following approximation:

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^1 \prod_{j=1}^{d} f_{j|X_1}(y_j|x)\, dx \approx \sum_{q=1}^{n_q} w_q \prod_{j=1}^{d} f_{j|X_1}(y_j|x_q),$$

where $\{x_q : q = 1, \ldots, n_q\}$ are the quadrature points and $\{w_q : q = 1, \ldots, n_q\}$ are the quadrature weights. To compute two-dimensional integrals for the 2-factor

model, the approximation uses Gauss-Legendre quadrature points in a double sum:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2 j|X_1}\big(x_2, y_j|x_1\big)\, dx_1 dx_2 \\ &\approx \sum_{q_1=1}^{n_q} \sum_{q_2=1}^{n_q} w_{q_1} w_{q_2} \prod_{j=1}^{d} f_{X_2 j|X_1}\big(x_{q_2}, y_j|x_{q_1}\big). \end{aligned}$$

With Gauss-Legendre quadrature, the same nodes and weights are used for different functions; this helps in yielding smooth numerical derivatives for numerical optimization via quasi-Newton (Nash, 1990). Our comparisons show that $n_q = 25$ quadrature points provide good precision.

2.3 Model selection

In this section we propose a heuristic method that automatically selects the bivariate parametric copula families that link the observed to the latent variables. This is very useful when the direction to the tail asymmetry based on semi-correlations is not consistent or clear. For multivariate mixed data, it is infeasible to estimate all possible combinations of bivariate parametric copula families, and compare them on the basis of information criteria. We develop an algorithm that can quickly select a factor copula model that accurately captures the (tail) dependence features in the data at hand. The linking copulas at each factor are selected with a sequential algorithm under the initial assumption that linking copulas are Frank, and then sequentially copulas with non-tail quadrant independence are assigned to any of pairs where necessary to account for tail asymmetry (discrete data) or tail dependence (continuous data).

For the 1-factor model, the proposed model selection algorithm is summarized in the following steps:

1. For $j = 1, \ldots, d$ estimate the marginal distributions $F_j(y)$.

2. Fit the 1-factor copula model with Frank copulas to link each of the d observed variables with the latent variable, i.e., maximise the log-likelihood function of the factor copula model in (2.5) over the vector of copula parameters $(\theta_1, \ldots, \theta_d)$.

3. If the jth linking copula has $\hat{\theta}_j > 0$, then select a set of copula candidates with ability to interpolate between independence and comonotonicity, otherwise select a set of copula candidates with ability to interpolate between countermonotonicity and independence.

4. For $j = 1, \ldots, d$:

   (a) fit all the possible 1-factor copula models, iterating over all the copula candidates for the jth variable;

   (b) select the copula family that corresponds to the lowest information criterion, say the Akaike, that is AIC $= -2 \times \ell + 2 \times \#$copula parameters;

   (c) fix the selected linking copula family for the jth variable.

For more than one factor we can select the appropriate linking copulas accordingly. We first select copula families in the first factor, and then we proceed to the next factor and apply exactly the same algorithm.

2.4 Techniques for parametric model comparison and goodness-of-fit

Factor copula models with different bivariate linking copulas can be compared via the log-likelihood or AIC at the maximum likelihood estimate. In addition, we will use the Vuong's test (Vuong, 1989) to show if a factor copula model provides better fit than the standard factor model with a latent additive structure, that is a factor copula model with BVN bivariate linking copulas (Krupskii and Joe, 2013; Nikoloulopoulos

and Joe, 2015). The Vuong's test is the sample version of the difference in Kullback-Leibler divergence between two models and can be used to differentiate two parametric models which could be non-nested. This test has been used extensively in the copula literature to compare vine copula models (e.g., Brechmann et al. 2012; Joe 2014; Nikoloulopoulos 2017). We provide specific details in Section 2.4.1.

Furthermore, to assess the overall goodness-of-fit of the factor copula models for mixed data, we will use appropriately the limited information M_2 statistic (Maydeu-Olivares and Joe, 2006). The M_2 statistic has been developed for goodness-of-fit testing in multidimensional contingency tables. Nikoloulopoulos and Joe (2015) has used the M_2 statistic to assess the goodness-of-fit of factor copula models for ordinal data. We build on the aforementioned papers and propose a methodology to assess the overall goodness-of-fit of factor copula models for mixed continuous and discrete responses. We provide the specifics for the M_2 statistic in Section 2.4.2.

2.4.1 Vuong's test for parametric model comparison

In this subsection, we summarize Vuong's test for comparing parametric models (Vuong, 1989). Assume that we have Models 1 and 2 with parametric densities $f^{(1)}_{\mathbf{Y}}$ and $f^{(2)}_{\mathbf{Y}}$, respectively. We can compare

$$\Delta_{1f_{\mathbf{Y}}} = n^{-1}\Big[\sum_{i=1}^{n}\Big\{E_{f_{\mathbf{Y}}}\log f_{\mathbf{Y}}(\mathbf{y}_i) - E_{f_{\mathbf{Y}}}\log f^{(1)}_{\mathbf{Y}}(\mathbf{y}_i;\boldsymbol{\theta}_1)\Big\}\Big],$$
$$\Delta_{2f_{\mathbf{Y}}} = n^{-1}\Big[\sum_{i=1}^{n}\Big\{E_{f_{\mathbf{Y}}}\log f_{\mathbf{Y}}(\mathbf{y}_i) - E_{f_{\mathbf{Y}}}\log f^{(2)}_{\mathbf{Y}}(\mathbf{y}_i;\boldsymbol{\theta}_2)\Big\}\Big].$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are the parameters in Models 1 and 2, respectively, that lead to the closest Kullback-Leibler divergence to the true $f_{\mathbf{Y}}$; equivalently, they are the limits in probability of the MLEs based on Models 1 and 2, respectively.

Model 1 is closer to the true $f_{\mathbf{Y}}$, i.e., is the better fitting model if $\Delta = \Delta_{1f_{\mathbf{Y}}} - \Delta_{2f_{\mathbf{Y}}} < 0$, and Model 2 is the better fitting model if $\Delta > 0$. The sample version of Δ with MLEs $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$ is

$$\bar{D} = \sum_{i=1}^{n} D_i/n,$$

where $D_i = \log\left[\frac{f_{\mathbf{Y}}^{(2)}(\mathbf{y}_i;\widehat{\boldsymbol{\theta}}_2)}{f_{\mathbf{Y}}^{(1)}(\mathbf{y}_i;\widehat{\boldsymbol{\theta}}_1)}\right]$. Vuong (1989) has shown that asymptotically that

$$\sqrt{n}\bar{D}/s \sim N(0,1),$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$. Hence, its 95% confidence interval (CI) is $\bar{D} \pm 1.96 \times \frac{1}{\sqrt{n}}\sigma$. In addition, Vuong (1989) provides adjusted log-likelihood ratios based on AIC correction (Joe, 2014). The AIC adjusted Vuong's 95% CI is

$$\bar{D} - n^{-1}[\dim(\widehat{\boldsymbol{\theta}}_2) - \dim(\widehat{\boldsymbol{\theta}}_1)] \pm 1.96 \times \frac{1}{\sqrt{n}}\sigma.$$

If the CIs include 0, then Model 1 and Model 2 are considered to be non-significantly different, while if the CIs are above 0, then Model 2 is favourable and considered to fit better than Model 1.

2.4.2 M_2 goodness-of-fit statistic

Since the M_2 statistic has been developed for multivariate ordinal data (Maydeu-Olivares and Joe, 2006), we propose to first transform the continuous and count variables to ordinal and then calculate the M_2 statistic at the maximum likelihood estimate before transformation.

Continuous variables can be transformed to ordinal with categories that are meaningful both practically and scientifically. If this is not the case, we propose an unsupervised strategy of transforming a continuous into an ordinal variable:

1. Set the number of ordinal categories K_j.
2. Transform Y_j to a standard uniform random variable U_j using its empirical distribution function.
3. Set the ordinal cutpoints on the uniform scale by generating a regular sequence from 1 to $K_j - 1$ and then dividing over K_j.
4. Divide the range of U_j into intervals with the ordinal cutpoints as breaks.
5. Transform U_j into an ordinal variable Y_j according to the interval in which its values fall.

Count variables that contain very high counts or very low counts, can be treated as ordinal where the first or the last category contains all the low or high counts, respectively, and their other values remain as they are. We further propose an unsupervised strategy of categorising a count into an ordinal variable:

1. Set the number of ordinal categories K_j.
2. Divide the range of Y_j into intervals with a regular sequence of length $K_j + 1$ from $\min(Y_j)$ to $\max(Y_j)$ as breaks.
3. Transform Y_j into an ordinal variable according to the interval in which its values fall.

After applying the transformations as above for each continuous or count variable, we have d ordinal variables $Y_1, \ldots, Y_d$ (both the original and the transformed ones) where the jth $(1 \leq j \leq d)$ variable consists of $K_j \geq 2$ categories labelled as $0, 1, \ldots, K_j - 1$. Consider the set of univariate and bivariate residuals that do not include category 0. This is a residual vector of dimension

$$s = \sum_{j=1}^{d} (K_j - 1) + \sum_{1 \leq j_1 < j_2 \leq d} (K_{j_1} - 1)(K_{j_2} - 1).$$

For a factor copula model with parameter vector $\boldsymbol{\theta}$ of dimension q, let $\boldsymbol{\pi}_2(\boldsymbol{\theta}) = \big(\dot{\boldsymbol{\pi}}_1(\boldsymbol{\theta})^\top, \dot{\boldsymbol{\pi}}_2(\boldsymbol{\theta})^\top\big)^\top$ be the column vector of the model-based marginal probabilities with $\dot{\boldsymbol{\pi}}_1(\boldsymbol{\theta})$ the vector of univariate marginal probabilities, and $\dot{\boldsymbol{\pi}}_2(\boldsymbol{\theta})$ the vector of bivariate marginal probabilities. Also, let $\mathbf{p}_2 = (\dot{\mathbf{p}}_1^\top, \dot{\mathbf{p}}_2^\top)^\top$ be the vector of the observed sample proportions, with $\dot{\mathbf{p}}_1$ the vector of univariate marginal proportions, and $\dot{\mathbf{p}}_2$ the vector of the bivariate marginal proportions.

With a sample size n, the limited information statistic M_2 is given by

$$M_2 = M_2(\hat{\boldsymbol{\theta}}) = n\big(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})\big)^\top \mathbf{C}_2(\hat{\boldsymbol{\theta}})\big(\mathbf{p}_2 - \boldsymbol{\pi}_2\big(\hat{\boldsymbol{\theta}}\big)\big), \tag{2.6}$$

with

$$\mathbf{C}_2(\boldsymbol{\theta}) = \boldsymbol{\Xi}_2^{-1} - \boldsymbol{\Xi}_2^{-1}\boldsymbol{\Delta}_2(\boldsymbol{\Delta}_2^\top\boldsymbol{\Xi}_2^{-1}\boldsymbol{\Delta}_2)^{-1}\boldsymbol{\Delta}_2^\top\boldsymbol{\Xi}_2^{-1} = \boldsymbol{\Delta}_2^{(c)}\big([\boldsymbol{\Delta}_2^{(c)}]^\top\boldsymbol{\Xi}_2\boldsymbol{\Delta}_2^{(c)}\big)^{-1}[\boldsymbol{\Delta}_2^{(c)}]^\top, \tag{2.7}$$

where $\boldsymbol{\Delta}_2 = \partial\boldsymbol{\pi}_2(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top$ is an $s \times q$ matrix with the derivatives of all the univariate and bivariate marginal probabilities with respect to the model parameters, $\boldsymbol{\Delta}_2^{(c)}$ is an $s \times (s-q)$ orthogonal complement to $\boldsymbol{\Delta}_2$, such that $[\boldsymbol{\Delta}_2^{(c)}]^\top\boldsymbol{\Delta}_2 = \mathbf{0}$, and $\boldsymbol{\Xi}_2 = \text{diag}(\boldsymbol{\pi}_2(\boldsymbol{\theta})) - \boldsymbol{\pi}_2(\boldsymbol{\theta})\boldsymbol{\pi}_2(\boldsymbol{\theta})^\top$ is the $s \times s$ covariance matrix of all the univariate and bivariate marginal sample proportions, excluding category 0. Due to equality in (2.7), $\mathbf{C}_2$ is invariant to the choice of orthogonal complement. The limited information statistic M_2 has a null asymptotic distribution that is χ^2 with $s-q$ degrees of freedom when the estimate $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent. For details on the computation of $\boldsymbol{\Xi}_2$ and $\boldsymbol{\Delta}_2$ for factor copula models we refer the interested reader to Nikoloulopoulos and Joe (2015).

2.5 Applications

In this section we illustrate the proposed methodology by re-analysing three mixed response datasets.

Initially, we use the diagnostic method in Joe (2014, pages 245-246) to show that each dataset (or more precisely the correlation matrix of the observed variables for each dataset) has a factor structure based on linear factor analysis. The correlation matrix $\mathbf{R}_{\text{observed}}$ has been obtained based on the sample correlations from the bivariate pairs of the observed variables. These are the linear (when both variables are continuous), polychoric (when both variables are discrete), and polyserial (when one variable is continuous and the other is discrete) sample correlations among the observed variables. The resulting $\mathbf{R}_{\text{observed}}$ is generally positive definite if the sample size is not small enough; if not one has to convert it to positive definite. We calculate various measures of discrepancy between $\mathbf{R}_{\text{observed}}$ and $\mathbf{R}_{\text{model}}$ (the resulting correlation matrix of linear factor analysis), such as the maximum absolute correlation difference $D_1 = \max|\mathbf{R}_{\text{model}} - \mathbf{R}_{\text{observed}}|$, the average absolute correlation difference $D_2 = \text{avg}|\mathbf{R}_{\text{model}} - \mathbf{R}_{\text{observed}}|$, and the correlation matrix discrepancy measure $D_3 = \log\big(\det(\mathbf{R}_{\text{model}})\big) - \log\big(\det(\mathbf{R}_{\text{observed}})\big) + \text{tr}(\mathbf{R}_{\text{model}}^{-1}\mathbf{R}_{\text{observed}}) - d$.

After confirming that a factor model with a parsimonious correlation structure is reasonable, we calculate the semi-correlations for each pair of observed variables to check if there is tail asymmetry. This will be a useful information for choosing potential parametric bivariate copulas other than the BVN copulas that lead to the standard factor model. Note that when the variables are negatively associated we calculate the sample semi-correlations in the lower-upper and upper-lower quadrant.

After motivating why more flexible dependencies are needed in cases of mixed data and how those dependencies in the data can be captured by suitable bivariate copulas, we proceed with factor copula models and construct a plausible factor cop-

ula model, to capture any type of reflection asymmetric dependence, by using the proposed algorithm in Section 2.3. For a baseline comparison, we first fit the factor copula models with the comprehensive bivariate parametric copula families that allow for reflection symmetric dependence; these are the BVN, Frank, and t_ν copulas. For t_ν copulas, we summarize the choice of integer ν with the largest log-likelihood. For the standard 2-factor model, to obtain a unique solution we must impose sufficient constraints. One parameter for the second factor can be set to zero and the likelihood can be maximized with respect to other $2d-1$ parameters. We report the varimax transform (Kaiser, 1958) of the loadings (a reparametrization of $2d$ parameters), converted to factor copula parameters via the relations

$$\theta_j = \beta_{j1}, \quad \delta_j = \frac{\beta_{j2}}{(1-\beta_{j1}^2)^{1/2}}, \tag{2.8}$$

where β_{j1} and β_{j2} are the loadings at the first and second factor, respectively (Krupskii and Joe, 2013; Nikoloulopoulos and Joe, 2015).

If the number of parameters is not the same between the models, we use the AIC as a rough diagnostic measure for goodness-of-fit between the models, otherwise we use the likelihood at the maximum likelihood estimates. We further compute the Vuong's tests with Model 1 being the factor copula model with BVN copulas, that is the standard factor model, to reveal if any other factor copula model provides better fit than the standard factor model. To make it easier to compare strengths of dependence, we convert the estimated parameters to Kendall's τ's in $(-1,1)$ via the relations in Joe (2014, Chapter 4); SEs are also converted via the delta method. For the model that provides the best fit, we provide the estimates and SEs that are obtained by maximizing the joint likelihood in (2.5) at one step over $\boldsymbol{\theta}$. Although, the two-stage estimation approach in Section 2.2 is a convenient way to quickly compare candidate factor copula models, the full likelihood is applied for the best fitting factor copula

model. The overall fit of the factor copula models is evaluated using the M_2 statistic. Note that the M_2 statistic in the case with $2d - 1$ copulas (one set to independence for the second factor) is computed with $\boldsymbol{\Delta}_2$ having one less column.

2.5.1 Political-economic dataset

Quinn (2004) considered measuring the (latent) political-economic risk of 62 countries, for the year 1987. The political-economic risk is defined as the country's risk in manipulating economic rules for its own and constituents' advantages (see e.g., North and Weingast 1989). Quinn (2004) used 5 mixed variables, namely the continuous variable 'black-market premium' in each country (used as a proxy for illegal economic activity), the continuous variable productivity as measured by 'real gross domestic product per worker' in 1985 international prices, the binary variable 'independence of the national judiciary' (1 if the judiciary is judged to be independent and 0 otherwise), and the ordinal variables measuring the 'lack of expropriation risk' and 'lack of corruption'. The dataset and its complete description can be found in Quinn (2004) or in the `R` package `MCMCpack` (Martin et al., 2011). Note that since the continuous variable black-market premium is negatively associated with the remaining variables (from the context), we re-orient it leading to positive dependence among all the observed variables.

Table 2.2 shows that the sample correlation matrix of the mixed responses has an 1-factor structure based on linear factor analysis (large D_3 is due to the small sample size as demonstrated using simulated data in Section 2.6). The sample semi-correlations in Table 2.2 show that there is more probability in the upper tail or lower tail compared with a discretized MVN, suggesting that a factor model with bivariate parametric copulas with upper or lower tail dependence might provide a better fit. Table 2.3 gives the estimated parameters, their standard errors (SE) in Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's tests, and the M_2 statistics for

Table 2.2: The sample correlation ρ_N, lower semi-correlation ρ_N^-, and upper semi-correlation ρ_N^+ for each pair of variables, along with the measures of discrepancy between the sample and the resulting correlation matrix of linear factor analysis with 1 and 2 factors for the political-economic risk data.

pairs of variables		ρ_N	ρ_N^-	ρ_N^+
BM	GDP	0.53	-0.04	0.57
BM	IJ	0.61	-	-
BM	XPR	0.67	0.88	0.63
BM	CRP	0.62	0.16	0.55
GDP	IJ	0.78	-	-
GDP	XPR	0.55	0.11	0.75
GDP	CRP	0.77	0.24	0.63
IJ	XPR	0.91	-	-
IJ	CRP	0.87	-	-
XPR	CRP	0.76	0.71	0.71

# factors	D_1	D_2	D_3
1	0.16	0.04	0.91
2	0.06	0.01	0.22

BM: black-market premium; CPR: lack of corruption; GDP: gross domestic product; IJ: independent judiciary; XPR: lack of expropriation risk.

the 1-factor copula models. Table 2.3 also indicates the parametric copula family chosen for each pair using the proposed heuristic algorithm. Copulas with asymmetric dependence are selected for all the copulas that link the latent variable to each of the observed variables. Hence, it is revealed that there are features in the data such as tail dependence and asymmetry which cannot be captured by copulas with reflection symmetric dependence such as BVN, Frank and t_ν copulas.

Table 2.3: Estimated parameters, their standard errors (SE) in Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's statistics, and the M_2 statistics for the one-factor copula models for the political-economic risk data.

1-factor	BVN[¶]		t_5		Frank		Selected		
	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\hat{\tau}$	SE	copulas	$\hat{\tau}$	SE
BM	0.50	0.06	0.51	0.07	0.49	0.06	Joe	0.51	0.05
GDP	0.57	0.05	0.57	0.06	0.58	0.06	Joe	0.58	0.05
IJ	0.80	0.09	0.81	0.09	0.75	0.09	reflected Joe	0.80	0.07
XPR	0.66	0.06	0.68	0.07	0.66	0.06	Joe	0.69	0.06
CRP	0.71	0.06	0.70	0.06	0.72	0.06	Gumbel	0.74	0.06
ℓ	-165.15		-166.25		-164.89		-151.98		
Vuong 95%CI			(-0.051,0.015)		(-0.077,0.085)		(0.073,0.352)		
M_2	179.2		187.4		177.6		129.2		
df	134		134		134		134		
p-value	< 0.01		< 0.01		< 0.01		0.60		

[¶]: The resulting model is the same as the standard factor model; BM: black-market premium; GDP: gross domestic product; IJ: independent judiciary; XPR: lack of expropriation risk; CPR: lack of corruption.

In all of the fitted models the estimated Kendall's τ's are similar. Kendall's τ only accounts for the dependence dominated by the middle of the data, and it is expected to be similar amongst different families of copulas. However, the tail dependence and tail order vary, as explained in Section 1.3, and they are properties to consider when choosing amongst different families of copulas (Nikoloulopoulos and Karlis, 2008).

The table shows that the selected model using the proposed algorithm provides the best fit and there is a substantial improvement over the standard factor model as indicated by the Vuong and M_2 statistics. To compute the M_2 statistics we transformed the continuous variables to ordinal with 5 categories using the unsupervised strategy in Section 2.4.2; a similar inference was drawn, when we transformed them to ordinal with 3, 4, or 6 categories. The factor copula parameter of 0.51 on negative black market premium indicates a negative association between the illegal economic activity and the latent variable. All the other estimated factor copula parameters indicate a positive association between each of the other observed variables (independent judiciary, productivity, lack of expropriation, and lack of corruption) with the latent variable. Hence, we can interpret the latent variable to be the political economical certainty.

2.5.2 General social survey

Hoff (2007) analysed seven demographic variables of 464 male respondents to the 1994 General Social Survey. Of these seven, two were continuous (income and age of the respondents), three were ordinal with 5 categories (highest degree of the survey respondent, income and highest degree of respondent's parents), and two were count variables (number of children of the survey respondent and respondent's parents). The data are available in Hoff (2007, Supplemental materials).

Table 2.4: The sample correlation ρ_N, lower semi-correlation ρ_N^-, and upper semi-correlation ρ_N^+ for each pair of variables, along with the measures of discrepancy between the sample and the resulting correlation matrix of linear factor analysis with 1, 2 and 3 factors for the general social survey dataset.

pairs of variables		ρ_N	ρ_N^-	ρ_N^+
income	age	0.29	0.48	0.23
income	degree	0.52	0.24	0.33
income	pincome	0.14	0.02	0.28
income	pdegree	0.24	0.04	0.08
income	child	0.22	0.23	0.01
income	pchild	-0.09	0.06	0.00
age	degree	0.06	0.22	-0.04
age	pincome	-0.11	-0.02	0.12
age	pdegree	-0.14	-0.42	0.44
age	child	0.58	0.36	0.26
age	pchild	0.12	0.18	0.07
degree	pincome	0.21	0.17	-0.05
degree	pdegree	0.46	0.46	0.41
degree	child	-0.11	-0.10	-0.09
degree	pchild	-0.25	-0.14	-0.30
pincome	pdegree	0.44	0.44	0.34
pincome	child	-0.16	-0.15	0.11
pincome	pchild	-0.23	0.13	-0.30
pdegree	child	-0.21	0.08	0.10
pdegree	pchild	-0.34	0.19	-0.32
child	pchild	0.20	-0.11	-0.06

# factors	D_1	D_2	D_3
1	0.55	0.09	0.82
2	0.15	0.03	0.13
3	0.02	0.00	0.00

Table 2.4 shows that the sample correlation matrix of the mixed responses has a 2- or even a 3-factor structure based on linear factor analysis. The direction of the tail asymmetry based on sample semi-correlations in Table 2.4 is not consistent, and this shows the usefulness of the proposed model selection technique. Table 2.5 gives the estimated parameters, their standard errors (SE) in Kendall's τ scale, the joint log-likelihoods, the 95% CIs of Vuong's tests, and the M_2 statistics for the 1-factor

and 2-factor copula models. The best fit for the 1-factor model is based on the bivariate copulas selected by the proposed algorithm, where there is improvement over the factor copula model with BVN copulas according to Vuong's statistic. However, assessing the overall goodness-of-fit via the M_2 statistic, it is revealed that one latent variable is not adequate to explain the dependencies among the mixed responses. To apply the M_2 statistic, age and income were transformed to ordinal with 4 (18–24, 25–44, 45–64, and $65+$) and 5 (0–10, 11–19, 20–29, 30–40, and $41+$) categories, respectively, and number of children of the survey respondent and respondent's parents were treated as ordinal where the 4th (more than 3 children) and 8th (more than 7 children) category, respectively, contained all the high counts.

The 2-factor copula models with BVN, t_ν, and Frank copulas provide some improvement over the 1-factor copula models but according to the M_2 statistic they still have a poor fit. Note that the factor copula model with t_9 copulas was not identifiable (large SEs) in line with Nikoloulopoulos and Joe (2015), hence one parameter for the second factor was set to zero and the likelihood was maximized with respect to the remaining parameters. We report the varimax transform (Kaiser, 1958) of the loadings, converted to factor copula parameters via the relations in (2.8).

The selected 2-factor copula model using the algorithm in Section 2.3 shows improvement over the standard factor model according to the Vuong's statistic and better fit according to the M_2 statistic; it changes a p-value < 0.001 to one > 0.10. For the 2-factor model based on the proposed algorithm for model selection, note that, without the need for a varimax rotation, the unique loading parameters ($\hat{\tau}$'s converted to normal copula parameters $\hat{\theta}_j$'s and $\hat{\delta}_j$'s and then to loadings using the relations in (2.8)) show that one factor is loaded only on the demographic variables of the respondent's parents.

Table 2.5: Estimated parameters, their standard errors (SE) in Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's statistics, and the M_2 statistics for the 1- and 2-factor copula models for the general social survey dataset.

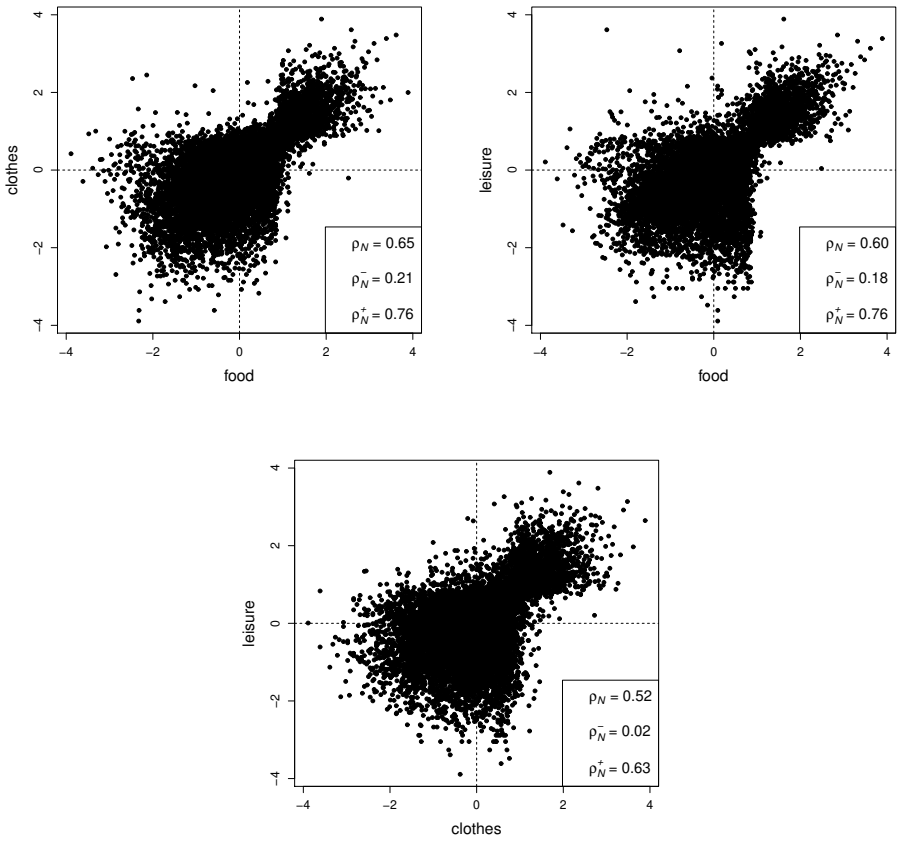
1-factor	BVN¶		t_9		Frank		Selected		
	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\hat{\tau}$	SE	copulas	$\hat{\tau}$	SE
income	0.20	0.04	0.20	0.04	0.20	0.04	Joe	0.29	0.04
age	-0.14	0.04	-0.14	0.04	-0.14	0.04	2-reflected Joe	-0.14	0.03
degree	0.40	0.04	0.39	0.04	0.38	0.04	t_3	0.45	0.04
pincome	0.33	0.03	0.34	0.04	0.35	0.04	t_3	0.33	0.05
pdegree	0.62	0.05	0.65	0.05	0.68	0.06	reflected Gumbel	0.56	0.05
child	-0.20	0.04	-0.19	0.04	-0.19	0.04	2-reflected Joe	-0.14	0.03
pchild	-0.32	0.03	-0.31	0.04	-0.32	0.04	2-reflected Gumbel	-0.27	0.03
ℓ	-3425.39		-3420.56		-3433.83		-3397.79		
Vuong 95%CI			(-0.005,-0.025)		(-0.037,0.001)		(0.022,0.097)		
M_2	743.74		715.45		738.76		660.47		
df	348		348		348		348		
p-value	< 0.001		< 0.001		< 0.001		< 0.001		
2-factor	BVN¶		t_9		Frank		Selected		
	$\hat{\tau}$		$\hat{\tau}$		$\hat{\tau}$	SE	copulas	$\hat{\tau}$	SE
1st factor									
income	0.36		0.35		0.13	0.04	reflected Gumbel	0.34	0.03
age	-0.05		-0.06		0.50	0.05	reflected Joe	0.49	0.03
degree	0.55		0.53		-0.12	0.04	BVN	0.18	0.04
pincome	0.27		0.28		-0.21	0.04	1-reflected Joe	-0.13	0.04
pdegree	0.48		0.50		-0.31	0.05	1-reflected Joe	-0.13	0.04
child	-0.13		-0.14		0.52	0.05	reflected Joe	0.44	0.04
pchild	-0.28		-0.28		0.23	0.04	Gumbel	0.11	0.03
2nd factor									
income	0.38		0.41		0.50	0.06	Gumbel	0.40	0.04
age	0.54		0.55		0.21	0.04	2-reflected Joe	-0.14	0.03
degree	0.14		0.17		0.57	0.07	reflected Joe	0.65	0.06
pincome	-0.09		-0.08		0.23	0.04	Gumbel	0.30	0.04
pdegree	-0.16		-0.14		0.44	0.05	t_5	0.49	0.04
child	0.53		0.53		0.08	0.04	BVN	-0.24	0.04
pchild	0.13		0.10		-0.24	0.04	2-reflected Gumbel	-0.26	0.03
ℓ	-3286.80		-3278.88		-3300.07		-3235.86		
Vuong 95%CI			(-0.004,-0.038)		(-0.058,0.001)		(0.061,0.159)		
M_2	471.47		461.70		492.37		370.61		
df	342		342		341		341		
p-value	< 0.001		< 0.001		< 0.001		0.13		

¶: The resulting model is the same as the standard factor model; pdemographic: demographic variable of respondent's parents.

2.5.3 Swiss consumption survey

Irincheeva et al. (2012b) considered measuring the latent variable 'financial wealth of the household' in its different realizations by analysing seven household variables of $n = 9960$ respondents to the Swiss consumption survey. Out of these seven, three were continuous (food, clothing and leisure expenses), three were binary (dishwasher, car, and motorcycle), and one was a count variable (the number of bicycles in possession of the household).

Figure 2.1: Bivariate normal scores plots, along with correlations and semi-correlations for the continuous data from the Swiss consumption survey.



Irincheeva et al. (2012b), with simple descriptive statistics such as scatter plots of the original data, have shown that these mixed responses have reflection asymmetric dependence, and fitted their latent variable approach with one and two latent variables. In Figure 2.1 we depict the bivariate normal scores plots for the continuous data along with their correlations and semi-correlations. With a bivariate normal scores plot one can check for deviations from the elliptical shape that would be expected with the BVN copula, and hence assess if tail asymmetry and tail dependence exists on the data. For all the pairs the upper semi-correlation is larger, and interestingly, contrasting the bivariate normal scores plots in Figure 2.1 with the contour plots in Figure 1.2, it is apparent that for the continuous variables the linking copulas might be the BB10 copulas.

Table 2.6 shows that the sample correlation matrix of the mixed responses has a 2-factor structure based on linear factor analysis. The sample semi-correlations in Table 2.6 show that there is more probability in the upper tail and lower tail among the continuous variables and between each of the continuous variables with the count variable, respectively, suggesting that a factor model with bivariate parametric copulas with asymmetric tail dependence might provide a better fit. Table 2.7 gives the estimated parameters, their standard errors (SE) in Kendall's tau scale, the joint log-likelihoods, the 95% CIs of Vuong's test, and the M_2 statistics for the 1-factor and 2-factor copula models. The best fitted 1- and 2-factor models result when we use BB10 copulas with asymmetric quadrant tail independence to link the latent variable to each of the continuous observed variables and copulas with lower tail dependence to link the latent variables to the discrete observed variables. Once again the one-factor copula model is not adequate to explain the dependence amongst the mixed responses based on the M_2 statistic (Table 2.7, 1-factor). To apply the M_2 statis-

Table 2.6: The sample correlation ρ_N, lower semi-correlation ρ_N^-, and upper semi-correlation ρ_N^+ for each pair of variables, along with the measures of discrepancy between the sample and the resulting correlation matrix of linear factor analysis with 1, 2 and 3 factors for the Swiss consumption survey dataset.

pairs of variables		ρ_N	ρ_N^-	ρ_N^+
food	clothes	0.65	0.21	0.76
food	leisure	0.60	0.18	0.76
food	dishwasher	0.31	-	-
food	car	0.38	-	-
food	motorcycle	0.11	-	-
food	bicycles	0.21	0.22	0.02
clothes	leisure	0.52	0.02	0.63
clothes	dishwasher	0.23	-	-
clothes	car	0.25	-	-
clothes	motorcycle	0.07	-	-
clothes	bicycles	0.18	0.15	0.02
leisure	dishwasher	0.24	-	-
leisure	car	0.18	-	-
leisure	motorcycle	0.01	-	-
leisure	bicycles	0.08	0.04	0.08
dishwasher	car	0.43	-	-
dishwasher	motorcycle	0.03	-	-
dishwasher	bicycles	0.24	-	-
car	motorcycle	0.18	-	-
car	bicycles	0.26	-	-
motorcycle	bicycles	0.21	-	-

# factors	D_1	D_2	D_3
1	0.27	0.06	0.26
2	0.12	0.02	0.06
3	0.03	0.01	0.01

tic, we transformed the continuous to ordinal variables with 3 categories using the unsupervised strategy in Section 2.4.2 and the count variable bicycle was treated as ordinal where the 6th category contained all the high counts (5 bicycles or more).

While it is revealed that the selected 2-factor copula model is the best model (lowest AIC) and there is substantial improvement over the standard 2-factor model, it is not apparent from the M_2 statistic that the response patterns are satisfactorily

Table 2.7: Estimated parameters, their standard errors (SE) in Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's statistics, and the M_2 statistics for the 1- and 2-factor copula models for the Swiss consumption survey dataset.

1-factor	BVN[¶]		t_5		Frank		Selected		
	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\hat{\tau}$	SE	Copulas	$\hat{\tau}$	SE
food	0.69	0.01	0.73	0.01	0.74	0.01	reflected BB10	0.79	0.00
clothes	0.53	0.01	0.53	0.01	0.53	0.01	BB10	0.38	0.00
leisure	0.47	0.01	0.50	0.01	0.50	0.01	BB10	0.39	0.00
dishwasher	0.24	0.01	0.25	0.01	0.23	0.01	reflected Joe	0.28	0.01
car	0.27	0.01	0.30	0.01	0.28	0.01	reflected Joe	0.23	0.01
motorcycle	0.07	0.01	0.06	0.01	0.08	0.01	reflected Joe	0.13	0.01
bicycles	0.15	0.01	0.15	0.01	0.16	0.01	reflected Joe	0.17	0.01
AIC	55004.24		54221.36		55105.88		48932.32		
Vuong 95% CI			(0.032,0.046)		(-0.015,0.005)		(0.286,0.324)		
M_2	2775.73		2734.05		2808.53		1626.54		
df	71		71		71		68		
$p-$value	< 0.001		< 0.001		< 0.001		< 0.001		

2-factor	BVN[¶]	t_7		Frank		Selected		
	$\hat{\tau}$	$\hat{\tau}$	SE	$\hat{\tau}$	SE	copulas	$\hat{\tau}$	SE
1st factor								
food	0.61	0.34	0.03	0.48	0.01	BB10	0.38	0.00
clothes	0.51	0.32	0.03	0.42	0.01	BB10	0.36	0.01
leisure	0.49	0.35	0.02	0.42	0.01	BB10	0.38	0.01
dishwasher	0.14	-0.07	0.03	0.08	0.01	reflected Joe	0.19	0.02
car	0.12	-0.13	0.03	0.07	0.01	reflected Joe	0.10	0.01
motorcycle	0.01	-0.10	0.02	-0.08	0.01	Frank	0.02	0.01
bicycles	0.07	-0.10	0.02	-0.05	0.01	Frank	0.04	0.01
2nd factor								
food	0.36	0.66	0.01	0.66	0.01	BB10	0.53	0.01
clothes	0.18	0.46	0.02	0.40	0.01	BVN	0.28	0.01
leisure	0.07	0.41	0.02	0.36	0.01	BB10	0.30	0.01
dishwasher	0.33	0.37	0.01	0.26	0.01	BVN	0.42	0.01
car	0.48	0.46	0.02	0.36	0.01	reflected Joe	0.35	0.01
motorcycle	0.19	0.15	0.01	0.21	0.02	reflected Joe	0.17	0.01
bicycles	0.27	0.27	0.01	0.31	0.01	reflected Gumbel	0.27	0.01
AIC	54245.91	53482.23		53514.75		46233.00		
Vuong 95% CI		(0.032,0.045)		(0.028,0.046)		(0.386,0.419)		
M_2	1920.27	1886.66		1945.07		450.32		
df	65	64		64		59		
p-value	< 0.001	< 0.001		< 0.001		< 0.001		

[¶]: The resulting model is the same as the standard factor model.

explained by even 2 latent variables. This is not surprising since one should expect discrepancies between the postulated parametric model and the population probabilities, when the sample size is sufficiently large (Maydeu-Olivares and Joe, 2014). In Table 2.8 we list the maximum deviations of observed and expected counts for each bivariate margin, that is, $D_{j_1 j_2} = n \max_{y_1,y_2} |p_{j_1,j_2,y_1,y_2} - \pi_{j_1,j_2,y_1,y_2}(\hat{\boldsymbol{\theta}})|$. From the table, it is revealed, that there is no misfit. The maximum discrepancy occurs between the continuous variables 'food' and 'leisure'. For this bivariate margin, the discrepancy of 509/9960 maximum occurs in the BVN factor copula model, while this drops to 133/9960 in the selected 2-factor copula model.

Table 2.8: Maximum deviations $D_{j_1 j_2}$ of observed and expected counts for each bivariate margin (j_1, j_2) for the 1- and 2-factor copula models for the Swiss consumption survey dataset.

	1-factor model				2-factor model			
D_{j_1,j_2}	BVN	t_5	Frank	Selected	BVN	t_7	Frank	Selected
$D_{1,2}$	347	317	303	167	349	311	270	40
$D_{1,3}$	511	468	456	183	**509**	460	428	**133**
$D_{1,4}$	158	177	163	70	159	185	161	56
$D_{1,5}$	231	189	223	119	233	181	230	60
$D_{1,6}$	87	117	88	60	87	130	72	12
$D_{1,7}$	78	92	79	88	78	110	89	81
$D_{2,3}$	442	418	431	69	433	403	393	54
$D_{2,4}$	59	80	84	145	38	56	64	86
$D_{2,5}$	96	107	107	201	60	47	93	36
$D_{2,6}$	18	3	18	27	19	15	29	39
$D_{2,7}$	51	76	60	83	49	91	52	61
$D_{3,4}$	182	146	141	196	253	216	168	83
$D_{3,5}$	82	105	106	191	59	13	83	61
$D_{3,6}$	59	58	69	71	13	23	27	45
$D_{3,7}$	62	54	64	103	65	67	69	59
$D_{4,5}$	289	276	286	223	66	74	207	2
$D_{4,6}$	9	5	11	29	133	138	100	96
$D_{4,7}$	82	81	81	88	28	20	46	54
$D_{5,6}$	111	123	111	77	15	22	19	20
$D_{5,7}$	101	96	95	68	33	25	40	64
$D_{6,7}$	70	74	70	61	80	96	87	52

For the selected 2-factor model based on the proposed algorithm, note that, without the need for a varimax rotation, the unique loadings show that one factor is loaded only on the discrete variables (dishwasher, car, motorcycle, and bicycles), while both factors are loaded on the continuous variables (food, clothes, and leisure). This reveals that the one latent variable which is only associated with the continuous variables measures the expenses, while the the other which is associated with all the mixed variables measures the possession.

2.6 Simulations

An extensive simulation study is conducted to (a) examine the performance of the diagnostics to show that the correlation matrix of the simulated variables has a factor structure, (b) check the small-sample efficiency of the sample versions of $\rho_N, \rho_N^+, \rho_N^-$, (c) gauge the small-sample efficiency of the proposed estimation method and investigate the misspecification of the bivariate pair-copulas, (d) examine the reliability of using the heuristic algorithm to select the correct bivariate linking copulas, and (e) study the small-sample performance of the M_2 statistic after transforming the continuous and count variables to ordinal.

We randomly generated samples of size $n = \{100, 300, 500\}$ from each selected one- and two-factor copula models in the three application examples in Section 2.5. We set the type of the variables, the univariate margins and the bivariate linking copulas, along with their univariate and dependence parameters to mimic the real data. The binary variables don't have tail asymmetries, hence parametric copulas are less distinguishable. Therefore instead of binary, we simulated from ordinal with 3 equally weighted categories.

Table 2.9 contains the simulated means and standard deviations (SD) of the discrepancy measures D_1, D_2 and D_3. The resultant summaries show that all the discrepancy measures correctly recognize both that the correlation structure has a factor structure and the number of factors. Among the discrepancy measures, D_2 has a good performance even for a small sample size ($n = 100$), while this is not the case for D_1 and D_3 which require larger sample sizes to successively determine the number of adequate factors.

To check the small-sample efficiency of the sample versions of ρ_N, ρ_N^+ and ρ_N^- we have generated 10^4 random samples of size $n = \{100, 300, 500\}$ from all the aforementioned bivariate copulas that join the distributions of two continuous variables, two ordinal variables, one continuous and one ordinal variable, one continuous and one count variable, one ordinal and one count, and two count variables with small ($\tau = 0.3$), moderate ($\tau = 0.5$) and strong dependence ($\tau = 0.7$). Representative results are shown in Table 2.10 for the Gumbel copula. Note that the count variable was treated as ordinal with 5 categories where the 5th category contained all the counts greater than 3. The resultant biases, root mean square errors (RMSEs), and standard deviations (SDs), scaled by n, show the estimation of the correlations and semi-correlations is highly efficient. Note in passing that because only part of the data are used in computing sample semi-correlations, their variability is larger than the correlations. However, if there is a consistent direction to the tail asymmetry based on semi-correlations, this is useful information for choosing potential bivariate parametric copulas.

Table 2.9: Small sample of sizes $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected factor copula models in Section 2.5 to assess the measures of discrepancy D_1, D_2, and D_3 between the observed and the resulting correlation matrix of linear factor analysis for 1, 2 and 3 factors, with resultant means and standard deviations (SD).

		D_1		D_2		D_3	
n	# factors	mean	SD	mean	SD	mean	SD
Political-economic dataset – 1-factor model							
100	1	0.061	0.027	0.016	0.006	0.101	0.071
	2	0.022	0.016	0.004	0.003	0.014	0.023
300	1	0.038	0.017	0.010	0.004	0.036	0.023
	2	0.011	0.008	0.002	0.002	0.004	0.005
500	1	0.033	0.014	0.009	0.003	0.024	0.015
	2	0.009	0.006	0.002	0.001	0.002	0.003
General social survey – 1-factor model							
100	1	0.178	0.048	0.048	0.010	0.192	0.074
	2	0.119	0.037	0.025	0.006	0.077	0.039
	3	0.066	0.030	0.010	0.004	0.021	0.016
300	1	0.104	0.028	0.028	0.006	0.062	0.023
	2	0.068	0.021	0.015	0.004	0.024	0.012
	3	0.036	0.017	0.006	0.002	0.006	0.005
500	1	0.081	0.022	0.022	0.004	0.038	0.014
	2	0.053	0.016	0.012	0.003	0.014	0.007
	3	0.028	0.013	0.005	0.002	0.004	0.003
Swiss consumption survey – 1-factor model							
100	1	0.223	0.059	0.059	0.011	0.291	0.101
	2	0.144	0.046	0.029	0.007	0.106	0.053
	3	0.077	0.035	0.011	0.004	0.028	0.022
300	1	0.162	0.044	0.045	0.007	0.156	0.044
	2	0.091	0.030	0.018	0.005	0.036	0.019
	3	0.044	0.021	0.007	0.003	0.009	0.007
500	1	0.150	0.039	0.041	0.006	0.130	0.032
	2	0.071	0.024	0.014	0.004	0.022	0.011
	3	0.034	0.016	0.005	0.002	0.005	0.004
General social survey – 2-factor model							
100	1	0.360	0.066	0.102	0.018	0.691	0.183
	2	0.117	0.042	0.027	0.007	0.118	0.059
	3	0.059	0.028	0.010	0.004	0.028	0.023
300	1	0.332	0.045	0.101	0.012	0.573	0.103
	2	0.066	0.023	0.017	0.004	0.042	0.021
	3	0.033	0.015	0.006	0.003	0.009	0.008
500	1	0.326	0.037	0.101	0.010	0.552	0.078
	2	0.052	0.017	0.014	0.004	0.027	0.014
	3	0.026	0.012	0.005	0.002	0.006	0.005
Swiss consumption survey – 2-factor model							
100	1	0.249	0.070	0.060	0.013	0.343	0.129
	2	0.130	0.047	0.026	0.007	0.111	0.056
	3	0.065	0.031	0.010	0.004	0.028	0.023
300	1	0.200	0.047	0.048	0.009	0.198	0.061
	2	0.075	0.028	0.017	0.004	0.040	0.020
	3	0.036	0.017	0.006	0.003	0.009	0.007
500	1	0.191	0.038	0.046	0.007	0.171	0.045
	2	0.059	0.021	0.014	0.004	0.026	0.013
	3	0.027	0.013	0.005	0.002	0.006	0.005

Table 2.10: Small sample of sizes $n = \{100, 300, 500\}$ simulations (10^4 replications) from the Gumbel copula with Kendall's $\tau = \{0.3, 0.5, 0.7\}$ for mixed continuous, ordinal, and count data with resultant biases, root mean square errors (RMSE) and standard deviations (SD), scaled by n, for the estimated correlation ρ_N, lower semi-correlation ρ_N^-, and upper semi-correlation ρ_N^+.

			(continuous, continuous)			(continuous, ordinal)			(continuous, count)			(ordinal, ordinal)			(ordinal, count)			(count, count)		
n	τ		ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+
100	0.3	True values	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46
		nBias	-1.09	-0.79	-4.18	-1.05	0.16	-9.25	1.62	1.40	-4.41	-0.55	2.35	-9.86	0.62	4.88	-8.25	2.03	9.54	-5.58
		nSD	8.57	18.10	16.83	9.03	18.64	16.71	9.23	16.54	20.46	9.31	18.95	18.03	9.37	17.08	21.78	9.55	14.73	24.24
		nRMSE	8.64	18.12	17.34	9.09	18.64	19.10	9.37	16.60	20.93	9.32	19.10	20.55	9.39	17.76	23.29	9.76	17.55	24.87
	0.5	True values	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67
		nBias	-0.99	-1.65	-3.98	-0.16	-0.38	-10.21	2.43	0.10	-3.78	0.26	3.93	-8.58	1.41	7.42	-8.18	2.80	14.88	-5.57
		nSD	5.77	15.73	11.72	6.26	15.67	12.41	6.19	14.49	14.93	6.30	15.91	13.52	6.35	14.71	16.73	6.34	12.18	17.35
		nRMSE	5.85	15.82	12.37	6.26	15.68	16.07	6.65	14.49	15.40	6.31	16.39	16.01	6.51	16.48	18.62	6.94	19.23	18.22
	0.7	True values	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85
		nBias	-0.71	-2.12	-2.74	0.78	-2.21	-8.77	2.16	-5.21	-0.28	0.55	4.34	-4.46	1.23	6.29	-4.75	2.12	13.76	-2.19
		nSD	2.71	10.76	5.99	3.02	10.84	7.29	2.80	10.74	8.40	3.07	10.48	7.77	3.03	10.24	10.91	2.94	7.26	9.42
		nRMSE	2.80	10.96	6.59	3.12	11.06	11.40	3.53	11.94	8.40	3.12	11.35	8.96	3.27	12.02	11.90	3.63	15.56	9.67
300	0.3	True values	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46
		nBias	-1.44	-1.48	-5.96	-2.88	1.14	-26.54	5.52	4.43	-12.35	-1.56	7.52	-28.59	2.04	14.61	-25.56	6.36	28.76	-16.44
		nSD	15.04	30.94	28.32	15.75	31.26	27.83	16.11	28.02	33.34	16.32	32.42	30.34	16.45	28.98	36.06	16.65	25.39	40.28
		nRMSE	15.11	30.97	28.94	16.01	31.28	38.46	17.03	28.37	35.55	16.39	33.29	41.69	16.58	32.46	44.20	17.82	38.36	43.50
	0.5	True values	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67
		nBias	-1.23	-2.48	-5.34	-0.77	-1.16	-30.78	7.39	-0.39	-11.03	0.64	11.81	-25.37	4.11	21.74	-25.71	8.39	44.61	-16.40

Continued

Table 2.10 – *Continued*

			(continuous, continuous)			(continuous, ordinal)			(continuous, count)			(ordinal, ordinal)			(ordinal, count)			(count, count)		
n	τ		ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+
		nSD	9.99	26.98	19.09	10.87	26.60	20.40	10.62	24.72	24.25	11.08	27.48	22.59	11.09	25.22	27.56	10.96	20.84	28.88
		nRMSE	10.06	27.09	19.82	10.90	26.63	36.93	12.94	24.72	26.64	11.10	29.91	33.97	11.82	33.30	37.69	13.80	49.24	33.22
	0.7	True values	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85
		nBias	-0.83	-2.93	-3.43	1.42	-7.89	-28.35	5.84	-18.52	-1.43	1.31	12.56	-13.92	3.27	17.32	-16.87	5.97	40.56	-7.09
		nSD	4.60	18.37	9.35	5.16	18.37	11.94	4.71	18.05	13.58	5.34	18.16	13.05	5.26	17.59	18.02	5.04	12.35	15.54
		nRMSE	4.68	18.61	9.96	5.35	19.99	30.76	7.50	25.86	13.66	5.50	22.08	19.08	6.20	24.68	24.69	7.81	42.40	17.08
500	0.3	True values	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46	0.46	0.16	0.46
		nBias	-1.37	-1.08	-7.04	-4.45	2.42	-44.04	9.65	7.93	-20.25	-2.25	12.60	-47.33	3.75	24.71	-42.32	10.96	48.14	-27.35
		nSD	19.06	39.98	36.96	19.95	39.89	35.47	20.49	35.93	42.97	20.75	41.68	39.18	21.00	37.65	46.91	21.35	32.71	52.79
		nRMSE	19.11	40.00	37.63	20.44	39.97	56.55	22.64	36.79	47.51	20.87	43.54	61.45	21.33	45.04	63.17	24.00	58.20	59.46
	0.5	True values	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67	0.70	0.36	0.67
		nBias	-1.11	-2.31	-6.27	-1.16	-1.38	-51.40	12.58	-0.39	-18.48	1.34	19.78	-41.78	7.16	36.42	-42.79	14.31	74.64	-27.23
		nSD	12.58	35.08	24.56	13.67	34.07	26.18	13.31	31.87	31.24	14.04	35.55	29.22	13.99	32.68	36.00	13.91	26.57	37.57
		nRMSE	12.63	35.16	25.35	13.72	34.10	57.68	18.31	31.88	36.29	14.10	40.69	50.99	15.71	48.93	55.92	19.95	79.23	46.40
	0.7	True values	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85	0.88	0.64	0.85
		nBias	-0.76	-2.83	-3.63	2.03	-13.45	-47.97	9.47	-31.71	-2.93	2.21	21.01	-22.91	5.43	28.29	-28.60	9.95	67.47	-11.82
		nSD	5.81	23.66	11.69	6.48	23.48	15.30	5.95	23.22	17.23	6.75	23.36	16.73	6.68	22.72	23.21	6.42	15.65	20.03
		nRMSE	5.86	23.82	12.24	6.79	27.06	50.35	11.18	39.30	17.48	7.10	31.42	28.37	8.61	36.29	36.84	11.84	69.26	23.26

Table 2.11 contains the resultant biases, RMSEs, and SDs, scaled by n, for the estimates obtained using the estimation approach in Section 2.2. The results show that the proposed estimation approach is highly efficient according to the simulated biases, SDs and RMSEs. We further investigated the misspecification of the bivariate pair-copulas by deriving the same statistics but from 1-factor model with BVN pair copulas, i.e. the standard 1-factor model. Once again, the simulated data are based on the selected 1-factor copula models in Section 2.5. Table 2.12 contains the resultant biases, RMSEs, and SDs, scaled by n. The results show that the Kendall's tau estimates are not robust to pair-copula misspecification if the true (simulated) factor copula model has different dependence in the middle of the data, e.g. when the BB10 copulas that can provide a non convex shape of dependence (see e.g., Figure 1.2) are used to specify the true factor copula model (Table 2.12, Swiss consumption survey). As we have already mentioned the Kendall's τ only accounts for the dependence dominated by the middle of the data, and it is expected to be similar among parametric families of copulas that provide a convex shape of dependence (Table 2.12, Political-economic dataset and general social survey).

Table 2.13 contains four common nominal levels of the M_2 statistic under the factor copula models for mixed data. We transformed the continuous and count variables to ordinal with $K = \{3, 4, 5\}$ and $K = \{3, 4\}$ categories, respectively, using the unsupervised strategies proposed in Section 2.4.2. We also transformed the count variables to ordinal with $K = 5$ categories by treating them as ordinal where the 5th category contained all the counts greater than 3. As the observed levels are close to nominal levels, it is demonstrated that the M_2 statistic remains reliable for mixed data and that the information loss under transformation to ordinal is minimal.

Table 2.11: Small sample of sizes $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected factor copula models in Section 2.5 with resultant biases, root mean square errors (RMSE) and standard deviations (SD), scaled by n, for the estimated parameters.

Political-economic dataset – 1-factor model																				
τ	0.51			0.58			0.80			0.68			0.74							
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500					
nBias	0.88	2.30	3.17	-1.36	-3.39	-4.87	0.75	-0.55	0.64	0.21	-0.27	0.19	0.29	2.57	0.73					
nSD	4.28	7.60	9.63	4.19	7.50	9.08	5.41	10.91	11.98	4.58	8.43	9.84	4.46	14.92	11.78					
nRMSE	4.37	7.95	10.13	4.40	8.23	10.31	5.47	10.92	12.00	4.59	8.44	9.84	4.47	15.13	11.80					
General social survey – 1-factor model																				
τ	0.30			-0.14			0.46			0.33			0.55			-0.14			-0.27		
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
nBias	-0.11	-0.86	-1.66	-0.06	-0.10	-0.19	0.29	0.17	0.53	0.19	0.26	0.27	0.72	0.89	0.94	-0.18	-0.37	-0.30	-0.09	-0.03	-0.03
nSD	7.46	12.41	16.01	6.55	11.12	14.00	8.53	13.76	17.97	8.33	14.07	17.75	9.45	14.63	18.92	6.89	11.89	15.05	7.75	12.90	16.54
nRMSE	7.46	12.44	16.10	6.55	11.12	14.00	8.53	13.76	17.98	8.33	14.07	17.75	9.47	14.65	18.94	6.89	11.90	15.05	7.75	12.90	16.54
Swiss consumption survey – 1-factor model																				
τ	0.69			0.38			0.39			0.28			0.23			0.13			0.17		
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
nBias	-15.95	-0.78	-0.04	-7.85	-1.57	-3.22	-8.03	-1.62	-3.22	0.08	0.09	0.26	0.10	0.06	-0.11	0.23	0.12	0.40	0.13	0.06	0.20
nSD	8.81	9.93	13.16	9.58	6.24	7.98	9.54	6.52	8.11	7.69	13.02	16.80	7.72	13.02	17.02	7.46	13.01	16.90	7.51	12.79	16.67
nRMSE	18.23	9.96	13.16	12.38	6.43	8.60	12.47	6.72	8.73	7.69	13.02	16.81	7.72	13.02	17.02	7.46	13.01	16.91	7.51	12.79	16.67

Continued

Table 2.11 – *Continued*

General social survey – 2-factor model					$n = 500$									
	1st factor							2nd factor						
τ	0.34	0.49	0.18	-0.13	-0.13	0.44	0.11	0.40	-0.14	0.65	0.29	0.49	-0.24	-0.26
nBias	1.18	-7.19	1.40	0.31	0.19	1.45	-0.44	-0.96	0.19	-0.05	0.22	2.59	-2.47	0.00
nSD	16.17	17.21	19.25	18.83	18.63	19.05	17.66	18.52	18.32	22.72	17.68	26.90	21.77	16.33
nRMSE	16.21	18.65	19.30	18.84	18.63	19.11	17.67	18.54	18.32	22.72	17.68	27.03	21.91	16.33
Swiss consumption survey – 2-factor model					$n = 500$									
	1st factor							2nd factor						
τ	0.34	0.36	0.38	0.19	0.09	0.02	0.04	0.53	0.28	0.30	0.42	0.35	0.17	0.27
nBias	-2.31	-1.60	-0.69	-3.01	-1.04	-0.54	2.89	-4.27	0.64	1.00	3.41	1.27	0.59	-4.15
nSD	7.43	13.67	16.12	27.11	25.31	21.27	21.31	20.37	17.98	19.05	21.20	20.89	19.41	21.55
nRMSE	7.78	13.77	16.14	27.27	25.33	21.28	21.51	20.82	17.99	19.08	21.47	20.93	19.42	21.95

Table 2.12: Small sample of sizes $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected 1-factor copula models in Section 2.5 with resultant biases, root mean square errors (RMSE) and standard deviations (SD), scaled by n, for the estimated parameters under an 1-factor copula model with BVN copulas, i.e. the standard factor model.

Political-economic dataset – 1-factor model																	
τ	0.51			0.58			0.80			0.68			0.74				
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500		
nBias	-0.35	-0.96	-1.56	-1.40	-3.90	-6.41	-1.29	-6.19	-10.54	0.51	0.89	1.40	-0.18	-1.08	-2.15		
nSD	5.24	9.16	11.57	4.95	8.57	11.13	6.03	9.52	12.14	4.60	7.91	10.01	4.42	7.49	9.69		
nRMSE	5.25	9.21	11.68	5.15	9.42	12.85	6.17	11.35	16.07	4.63	7.96	10.11	4.42	7.57	9.92		
General social survey – 1-factor model																	
τ	0.30			-0.14			0.46			0.33			0.55			-0.14			-0.27		
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
nBias	-1.68	-5.15	-8.75	-0.93	-3.17	-5.45	-0.09	-1.34	-2.11	-0.16	-0.92	-1.86	0.65	-0.07	-0.75	-2.20	-6.94	-11.32	-0.99	-2.53	-4.30
nSD	7.66	12.91	16.57	8.14	13.62	17.45	9.08	14.45	18.82	8.56	14.28	18.05	10.46	15.79	20.38	8.67	14.79	18.91	8.51	13.60	17.53
nRMSE	7.84	13.90	18.73	8.19	13.99	18.28	9.08	14.52	18.94	8.57	14.31	18.15	10.48	15.79	20.39	8.95	16.34	22.03	8.57	13.83	18.05
Swiss consumption survey – 1-factor model																	
τ	0.69			0.38			0.39			0.28			0.23			0.13			0.17		
n	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
nBias	-16.40	-53.51	-90.99	3.02	8.67	14.58	3.02	8.30	13.91	-2.90	-8.03	-13.20	-2.26	-6.53	-11.09	-1.33	-4.02	-6.20	-3.02	-8.83	-14.50
nSD	12.86	21.36	26.83	10.08	17.97	23.59	10.36	18.07	23.67	9.36	16.40	21.17	9.17	15.68	20.71	8.77	15.06	19.80	8.36	14.33	18.63
nRMSE	20.84	57.62	94.87	10.53	19.95	27.73	10.79	19.88	27.45	9.80	18.27	24.94	9.45	16.99	23.49	8.87	15.58	20.75	8.88	16.83	23.61

Table 2.13: Small sample of sizes $n = \{100, 300, 500\}$ distribution for M_2 (10^4 replications). Empirical rejection levels at $\alpha = \{0.20, 0.10, 0.05, 0.01\}$, degrees of freedom (df), and mean under the factor copula models. Continuous and count variables are transformed to ordinal with $K = \{3, 4, 5\}$ and $K = \{3, 4\}$ categories, respectively, using the general strategies proposed in Section 2.4.2. Count variables area also transformed to ordinal with $K = 5$ categories by treating them as ordinal where the 5th category contained all the counts greater than 3.

	$n = 100$			$n = 300$			$n = 500$		
	$K = 3$	$K = 4$	$K = 5$	$K = 3$	$K = 4$	$K = 5$	$K = 3$	$K = 4$	$K = 5$
Political-economic dataset – 1-factor model									
df	92	121	152	92	121	152	92	121	152
mean	89.3	118.3	148.4	91.0	119.7	152.6	91.0	119.6	152.3
$\alpha = 0.20$	0.183	0.192	0.197	0.196	0.194	0.195	0.196	0.189	0.190
$\alpha = 0.10$	0.121	0.125	0.134	0.122	0.121	0.119	0.114	0.109	0.109
$\alpha = 0.05$	0.083	0.089	0.098	0.076	0.077	0.077	0.072	0.070	0.067
$\alpha = 0.01$	0.044	0.046	0.055	0.036	0.034	0.037	0.027	0.030	0.026
General social survey – 1-factor model									
df	161	239	329	161	239	329	161	239	329
mean	161.5	240.0	333.0	160.7	239.4	329.7	161.3	240.2	329.6
$\alpha = 0.20$	0.213	0.220	0.240	0.202	0.216	0.203	0.211	0.228	0.212
$\alpha = 0.10$	0.110	0.121	0.122	0.106	0.118	0.102	0.118	0.127	0.108
$\alpha = 0.05$	0.058	0.070	0.061	0.054	0.067	0.051	0.065	0.073	0.056
$\alpha = 0.01$	0.013	0.018	0.014	0.014	0.019	0.012	0.016	0.023	0.011
Swiss consumption survey – 1-factor model									
df	74	128	194	74	128	194	74	128	194
mean	75.4	130.1	197.8	74.6	128.5	195.1	74.5	128.0	194.4
$\alpha = 0.20$	0.229	0.239	0.254	0.214	0.209	0.221	0.210	0.202	0.207
$\alpha = 0.10$	0.121	0.135	0.147	0.111	0.104	0.113	0.105	0.099	0.103
$\alpha = 0.05$	0.067	0.076	0.086	0.056	0.055	0.060	0.051	0.053	0.053
$\alpha = 0.01$	0.016	0.024	0.030	0.011	0.013	0.013	0.012	0.011	0.012

	General social survey – 2-factor			Swiss consumption survey – 2-factor		
	$n = 500$			$n = 500$		
	$K = 3$	$K = 4$	$K = 5$	$K = 3$	$K = 4$	$K = 5$
df	154	232	322	65	119	185
mean	154.8	234.0	323.3	65.6	119.7	185.5
$\alpha = 0.20$	0.217	0.234	0.214	0.217	0.215	0.217
$\alpha = 0.10$	0.113	0.131	0.116	0.114	0.111	0.113
$\alpha = 0.05$	0.065	0.075	0.059	0.060	0.057	0.060
$\alpha = 0.01$	0.018	0.022	0.018	0.013	0.013	0.017

Table 2.14 presents the number of times that the true bivariate parametric copulas are chosen over 100 simulation runs. If the true copula has distinct dependence properties with medium to strong dependence, then the algorithm performs extremely

well as the sample size increases. Low selection rates occur if the true copulas have low dependence or similar tail dependence properties, since for that case it is difficult to distinguish amongst parametric families of copulas (Nikoloulopoulos and Karlis, 2008). For example,

Table 2.14: Frequencies of the true bivariate copula identified using the model selection algorithm from 100 simulation runs. Note: rCopula: reflected copula; 1rCopula: 1-reflected copula; 2rCopula: 2-reflected copula.

Political-economic dataset – 1-factor model						
	Continuous		Ordinal			
n	1rJoe	Joe	rJoe	Joe	Gumbel	
100	88	81	45	82	34	
300	88	93	54	83	60	
500	91	100	66	100	79	
General social survey – 1-factor model						
	Continuous		Ordinal			Count	
n	Joe	2rJoe	t_5	t_5	rGumbel	2rJoe	2r Gumbel
100	68	63	27	19	27	56	28
300	89	79	41	43	49	65	55
500	91	85	61	65	74	73	68
Swiss consumption survey – 1-factor model							
	Continuous			Ordinal			Count
n	rBB10	BB10	BB10	rJoe	rJoe	rJoe	rJoe
100	27	94	91	61	60	41	56
300	50	99	98	64	71	63	68
500	70	98	98	68	74	71	72
General social survey – 2-factor model							
1st Factor	Continuous		Ordinal			Count	
n	rGumbel	rJoe	BVN	1rJoe	1rJoe	rJoe	Gumbel
100	22	40	10	19	19	50	6
300	26	52	11	42	36	79	16
500	19	67	13	52	53	83	39
2nd Factor	Continuous		Ordinal			Count	
n	Gumbel	2rJoe	rJoe	Gumbel	t_5	BVN	2rGumbel
100	13	28	28	7	14	21	17
300	26	39	56	30	45	28	47
500	32	67	65	53	59	33	70
Swiss consumption survey – 2-factor model							
1st Factor	Continuous			Ordinal			Count
n	BB10	BB10	BB10	rJoe	rJoe	Frank	Frank
100	57	77	55	31	28	23	34
300	81	94	82	51	40	19	21
500	88	94	87	49	50	21	16
2nd Factor	Continuous			Ordinal			Count
n	BB10	BVN	BB10	BVN	rJoe	rJoe	rGumbel
100	5	14	28	10	29	31	10
300	27	29	43	22	49	40	16
500	39	39	60	31	55	63	31

- in the results from the 2-factor model for the general social survey, the true copula for the first continuous variable (1st factor) is the reflected Gumbel with $\tau = 0.34$ and is only selected a considerable small number of times. The algorithm instead selected with a high probability the reflected Joe (results not shown here), because both reflected Joe and Gumbel copulas provide similar dependence properties, i.e., lower tail dependence.

- in the results from the 2-factor model for the Swiss consumption survey, the variables with Frank copulas have the lowest selection rates. This is due to the fact that their true Kendall's τ's parameters are close to 0 (independence).

2.7 Software

Our modelling framework is implemented in the package **FactorCopula** (Kadhem and Nikoloulopoulos, 2021c) within the open source statistical environment R (R Core Team, 2020). All the analyses presented in Sections 2.5.1 and 2.5.2 are given as code examples in the package. The manual of the package is provided as an appendix.

2.8 Chapter summary

We have extended the factor copula models proposed in Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015) to the case of mixed continuous and discrete responses. We have shown that the factor copula models (obtained from the proposed model selection algorithm) provide a substantial improvement over the standard factor models on the basis of the log-likelihood principle, Vuong's, and M_2 goodness-of-fit statistics. This improvement relies on the fact that the latent variable distribution is expressed via factor copulas instead of the MVN distribution.

Chapter 3

Structured factor copula models for item response data

In this chapter, we propose copula extensions for bi-factor and second-order models for items that can be split into non-overlapping groups, where each group of items has homogeneous dependence. They subsume the factor copula models proposed in Chapter 2 as special cases when all variables are discrete and arise from the same group. The construction of the bi-factor copula model exploits bivariate copulas to link the observed variables with the common and group-specific factors. While for the second-order copula model, there are bivariate copulas that link the observed variables to the group-specific factors, and also bivariate copulas that link the group-specific factors to the second-order factor.

To build plausible models, we propose model selection algorithms that automatically select suitable bivariate copulas for the bi-factor and second-order copula models for item response data. In order to evaluate the fit of the models, we also propose goodness-of-fit testing based on the M_2 statistic of Maydeu-Olivares and Joe (2006).

We examine the performance and reliability of the model selection algorithms and goodness-of-fit statistic in an extensive simulation study.

The bi-factor and second-order copula models are suitable for capturing different dependencies between and within different groups of observed variables, while allowing for tail asymmetry or more probability in the tails than would be expected with the MVN. We illustrate our methodology by re-analysing a real dataset, and show that the proposed models with linking copulas (selected by the model selection algorithm) provide better fit than the Gaussian bi-factor and second-order models.

The chapter is organised as follows. Section 3.1 introduces the bi-factor and second-order copula models for item response and discusses their relationship with the existing models. Estimation techniques and computational details are provided in Section 3.2. Section 3.3 proposes a heuristic method to select suitable bivariate copulas and build bi-factor and second-order copula models. Section 3.4 summarizes the assessment of goodness-of-fit of these models using the M_2 statistic of Maydeu-Olivares and Joe (2006). Section 3.5 contains an extensive simulation study to gauge the small-sample efficiency of the proposed estimation, investigate the misspecification of the bivariate copulas, and examine the reliability of the model selection and goodness-of-fit techniques. Section 3.6 presents an application of our methodology to real world data. We conclude with a summary in Section 3.8.

3.1 Bi-factor and second-order copula models

Let $\underbrace{Y_{11}, \ldots, Y_{d_1 1}}_{1}, \ldots, \underbrace{Y_{1g}, \ldots, Y_{d_g g}}_{g}, \ldots, \underbrace{Y_{1G}, \ldots, Y_{d_G G}}_{G}$ denote the item response variables classified into the G non-overlapping groups. There are d_g items in group

g; $g = 1, \ldots, G$, $j = 1, \ldots, d_g$ and collectively there are $d = \sum_{g=1}^{G} d_g$ items, which are all measured on an ordinal scale; $Y_{jg} \in \{0, \ldots, K_{jg} - 1\}$. Let the cutpoints in the uniform $U(0, 1)$ scale for the jg'th item be $a_{jg,k}$, $k = 1, \ldots, K - 1$, with $a_{jg,0} = 0$ and $a_{jg,K} = 1$. These correspond to $a_{jg,k} = \Phi(\alpha_{jg,k})$, where $\alpha_{jg,k}$ are cutpoints in the normal $N(0, 1)$ scale.

The bi-factor and second-order factor copula models are presented in Subsections 3.1.1 and 3.1.2, respectively. Subsection 3.1.3 discusses their relationship with the existing Gaussian bi-factor and second-order models.

3.1.1 Bi-factor copula model

Consider a common factor X_0 and G group-specific factors $X_1, \ldots, X_G$, where $X_0, X_1, \ldots, X_G$ are independent and standard uniformly distributed. Let Y_{jg} be the jth observed variable in group g, with y_{jg} being the realization. The bi-factor model assumes that $Y_{1g}, \ldots, Y_{d_g g}$ are conditionally independent given X_0 and X_g, and that Y_{jg} in group g does not depend on $X_{g'}$ for $g \neq g'$. Figure 3.1 depicts a graphical representation of the model.

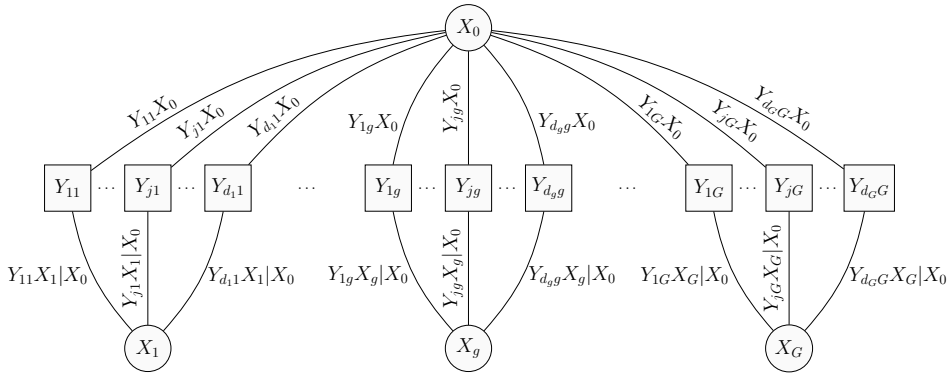


Figure 3.1: Graphical representation of the bi-factor copula model with G group-specific factors and a common factor X_0.

The joint probability mass function (pmf) is given by

$$\begin{aligned}\pi(\mathbf{y}) &= \Pr(Y_{jg} = y_{jg}; j = 1, \ldots, d_g, g = 1, \ldots, G) \\ &= \int_{[0,1]^{G+1}} \prod_{g=1}^{G} \prod_{j=1}^{d_g} \Pr(Y_{jg} = y_{jg} | X_0 = x_0, X_g = x_g) dx_1 \cdots dx_G dx_0.\end{aligned}$$

According to Sklar's theorem (Sklar, 1959) there exists a bivariate copula C_{Y_{jg},X_0} such that $\Pr(Y_{jg} \leq y_{jg}, X_0 \leq x_0) = C_{Y_{jg},X_0}\big(F_{Y_{jg}}(y_{jg}), x_0\big)$, for $x_0 \in [0,1]$, where C_{Y_{jg},X_0} is the copula that links observed variable with the common factor X_0, $F_{Y_{jg}}$ is the cumulative distribution function (cdf) of Y_{jg}; note that $F_{Y_{jg}}$ is a step function with jumps at $0, \ldots, K-1$, i.e., $F_{Y_{jg}}(y_{jg}) = a_{jg,y_{jg}+1}$. Then it follows that,

$$F_{Y_{jg}|X_0}(y_{jg}|x_0) := \Pr(Y_{jg} \leq y_{jg} | X_0 = x_0) = \frac{\partial}{\partial x_0} C_{Y_{jg},X_0}\big(F_{Y_{jg}}(y_{jg}), x_0\big).$$

For shorthand notation, we let $C_{Y_{jg}|X_0}\big(F_{Y_{jg}}(y_{jg})|x_0\big) = \frac{\partial}{\partial x_0} C_{Y_{jg},X_0}\big(F_{Y_{jg}}(y_{jg}), x_0\big)$.

The observed variables also load on the group-specific factors, hence to account for this dependence, we let $C_{Y_{jg},X_g|X_0}$ be a bivariate copula that links the observed variable $Y_{jg}$with the group-specific factor X_g given the common factor X_0. Hence,

$$\begin{aligned}&\Pr(Y_{jg} \leq y_{jg} | X_0 = x_0, X_g = x_g) = \frac{\partial}{\partial x_g} \Pr(Y_{jg} \leq y_{jg}, X_g \leq x_g | X_0 = x_0) \\ &= \frac{\partial}{\partial x_g} C_{Y_{jg},X_g|X_0}\big(F_{Y_{jg}|X_0}(y_{jg}|x_0), x_g\big) = C_{Y_{jg}|X_g;X_0}\big(F_{Y_{jg}|X_0}(y_{jg}|x_0)|x_g\big).\end{aligned}$$

To this end, the pmf of the bi-factor copula model takes the form

$$\begin{aligned}
\pi(\mathbf{y}) &= \int_{[0,1]^{G+1}} \prod_{g=1}^{G}\prod_{j=1}^{d_g}\Big\{ C_{Y_{jg}|X_g;X_0}\big(F_{Y_{jg}|X_0}(y_{jg}|x_0)|x_g\big) \\
&\quad - C_{Y_{jg}|X_g;X_0}\big(F_{Y_{jg}|X_0}(y_{jg}-1|x_0)|x_g\big)\Big\} dx_1\cdots dx_G dx_0 \\
&= \int_0^1 \prod_{g=1}^{G}\Bigg\{ \int_0^1 \prod_{j=1}^{d_g}\Big[C_{Y_{jg}|X_g;X_0}\big(F_{Y_{jg}|X_0}(y_{jg}|x_0)|x_g\big) \\
&\quad - C_{Y_{jg}|X_g;X_0}\big(F_{Y_{jg}|X_0}(y_{jg}-1|x_0)|x_g\big)\Big] dx_g\Bigg\} dx_0 \\
&= \int_0^1 \prod_{g=1}^{G}\Bigg\{ \int_0^1 \prod_{j=1}^{d_g}\Big[C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y_{jg}+1}|x_0)|x_g\big) \\
&\quad - C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y_{jg}}|x_0)|x_g\big)\Big] dx_g\Bigg\} dx_0 \\
&= \int_0^1 \prod_{g=1}^{G}\Bigg\{ \int_0^1 \prod_{j=1}^{d_g} f_{Y_{jg}|X_g;X_0}(y_{jg}|x_g,x_0) dx_g\Bigg\} dx_0. \qquad (3.1)
\end{aligned}$$

It is shown that the pmf is represented as an one-dimensional integral of a function which is in turn is a product of G one-dimensional integrals. Thus we avoid $(G+1)$-dimensional numerical integration.

For the parametric version of the bi-factor copula model, we let C_{Y_{jg},X_0} and $C_{Y_{jg},X_g|X_0}$ be parametric copulas with dependence parameters θ_{jg} and δ_{jg}, respectively.

3.1.2 Second-order copula model

Assume that for a fixed $g = 1,\ldots,G$, the items $Y_{1g},\ldots,Y_{d_g g}$ are conditionally independent given the first-order factors $X_g \sim U(0,1)$, $g = 1,\ldots,G$ and that $\mathbf{X} = (X_1,\cdots,X_G)$ are conditionally independent given the second-order factor $X_0 \sim U(0,1)$. That is the joint distribution of $\mathbf{X}$ has an one-factor structure. We also

assume that Y_{jg} in group g does not depend on $X_{g'}$ for $g \neq g'$. Figure 3.2 depicts the graphical representation of the model.

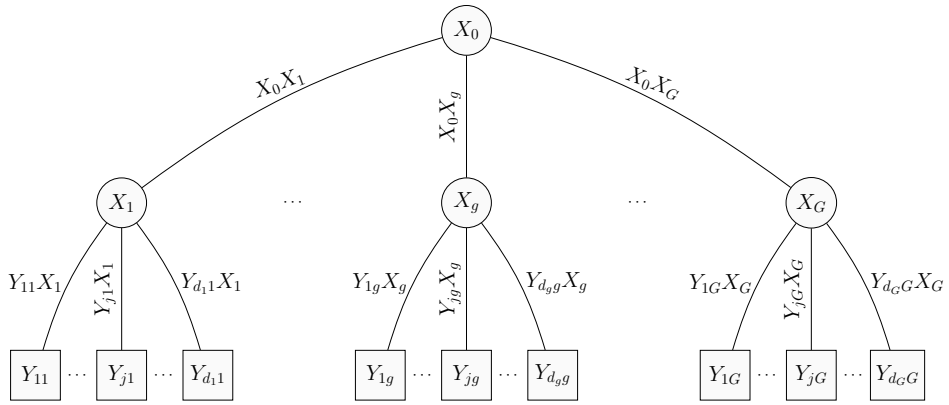


Figure 3.2: Graphical representation of the second-order copula model with G first-order factors and one second-order factor X_0.

The joint pmf takes the form

$$\pi(\mathbf{y}) = \int_{[0,1]^G} \left\{ \prod_{g=1}^{G} \prod_{j=1}^{d_g} \Pr(Y_{jg} = y_{jg} | X_g = x_g) \right\} c_{\mathbf{X}}(x_1, \ldots, x_G) dx_1 \cdots dx_G;$$

$c_{\mathbf{X}}$ is the one-factor copula density (Krupskii and Joe, 2013) of $\mathbf{X} = (X_1, \ldots, X_G)$, viz.

$$c_{\mathbf{X}}(x_1, \ldots, x_G) = \int_0^1 \prod_{g=1}^{G} c_{X_g, X_0}(x_g, x_0) dx_0,$$

where c_{X_g,X_0} is the bivariate copula density of the copula C_{X_g,X_0} linking X_g and X_0.

Letting C_{Y_{jg},X_g} be a bivariate copula that joins the observed variable Y_{jg} and the group-specific factor X_g such that

$$\begin{aligned} F_{Y_{jg}|X_g}(y_{jg}|x_g) :=& \Pr(Y_{jg} \leq y_{jg} | X_g = x_g) = \frac{\partial}{\partial x_g} C_{Y_{jg},X_g}\big(F_{Y_{jg}}(y_{jg}), x_g\big) \\ =& C_{Y_{jg}|X_g}\big(F_{Y_{jg}}(y_{jg})|x_g\big), \end{aligned}$$

the pmf of the second-order copula model becomes

$$\begin{aligned}
\pi(\mathbf{y}) &= \int_0^1 \int_{[0,1]^G} \left\{ \prod_{g=1}^{G} \prod_{j=1}^{d_g} \Big(C_{Y_{jg}|X_g}\big(F_{Y_{jg}}(y_{jg})|x_g\big) \right. \\
&\quad \left. - C_{Y_{jg}|X_g}\big(F_{Y_{jg}}(y_{jg}-1)|x_g\big) \Big) \right\} \left\{ \prod_{g=1}^{G} c_{X_g,X_0}\big(x_g,x_0\big) \right\} dx_1 \cdots dx_G dx_0 \\
&= \int_0^1 \left\{ \prod_{g=1}^{G} \int_0^1 \left[\prod_{j=1}^{d_g} \Big(C_{Y_{jg}|X_g}\big(F_{Y_{jg}}(y_{jg})|x_g\big) \right. \right. \\
&\quad \left. \left. - C_{Y_{jg}|X_g}\big(F_{Y_{jg}}(y_{jg}-1)|x_g\big) \Big) \right] c_{X_g,X_0}\big(x_g,x_0\big) dx_g \right\} dx_0 \\
&= \int_0^1 \left\{ \prod_{g=1}^{G} \int_0^1 \left[\prod_{j=1}^{d_g} \Big(C_{Y_{jg}|X_g}\big(a_{jg,y_{jg}+1}|x_g\big) \right. \right. \\
&\quad \left. \left. - C_{Y_{jg}|X_g}\big(a_{jg,y_{jg}}|x_g\big) \Big) \right] c_{X_g,X_0}\big(x_g,x_0\big) dx_g \right\} dx_0 \\
&= \int_0^1 \left\{ \prod_{g=1}^{G} \int_0^1 \Big[\prod_{j=1}^{d_g} f_{Y_{jg}|X_g}(y_{jg}|x_g) \Big] c_{X_g,X_0}\big(x_g,x_0\big) dx_g \right\} dx_0. \qquad (3.2)
\end{aligned}$$

Similarly with the bi-factor copula model, the pmf is represented as an one-dimensional integral of a function which is in turn is a product of G one-dimensional integrals.

For the parametric version of the second-order copula model, we let C_{Y_{jg},X_g} and C_{X_g,X_0} be parametric copulas with dependence parameters θ_{jg} and δ_g, respectively.

3.1.3 Special cases

In this subsection we show what happens when all bivariate copulas are BVN. Let Z_{jg} be the underlying continuous variable of the ordinal variable Y_{jg}, i.e., $Y_{jg} = y_{jg}$ if $\alpha_{jg,y_{jg}} \leq Z_{jg} \leq \alpha_{jg,y_{jg}+1}$ with $\alpha_{jg,K} = \infty$ and $\alpha_{jg,0} = -\infty$.

For the bi-factor model, and let $C_{Y_{jg}|X_g;X_0} = C_{Y_{jg}|X_g;X_0}(C_{Y_{jg}|X_0}(F_{jg}(y_{jg})|x_0)|x_g)$ for notational ease, if $C_{Y_{jg},X_0}(\cdot;\theta_{jg})$ and $C_{Y_{jg},X_g|X_0}(\cdot;\delta_{jg})$ are BVN copulas,

$$C_{Y_{jg}|X_g;X_0} = \Phi\left(\frac{\alpha_{jg,y_{jg}+1} - \theta_{jg}\Phi^{-1}(x_0) - \delta_{jg}\sqrt{1-\theta_{jg}^2}\Phi^{-1}(x_g)}{\sqrt{(1-\theta_{jg}^2)(1-\delta_{jg}^2)}}\right),$$

Hence, the pmf for the bi-factor copula model in (3.1) becomes

$$\begin{aligned}
\pi(\mathbf{y}) &= \int_0^1 \prod_{g=1}^{G}\left\{\int_0^1 \prod_{j=1}^{d_g}\left[\Phi\left(\frac{\alpha_{jg,y_{jg}+1} - \theta_{jg}\Phi^{-1}(x_0) - \delta_{jg}\sqrt{1-\theta_{jg}^2}\Phi^{-1}(x_g)}{\sqrt{(1-\theta_{jg}^2)(1-\delta_{jg}^2)}}\right) - \right.\right. \\
&\quad \left.\left.\Phi\left(\frac{\alpha_{jg,y_{jg}} - \theta_{jg}\Phi^{-1}(x_0) - \delta_{jg}\sqrt{1-\theta_{jg}^2}\Phi^{-1}(x_g)}{\sqrt{(1-\theta_{jg}^2)(1-\delta_{jg}^2)}}\right)\right] dx_g\right\} dx_0 \\
&= \int_{-\infty}^{\infty} \prod_{g=1}^{G}\left\{\int_{-\infty}^{\infty} \prod_{j=1}^{d_g}\left[\Phi\left(\frac{\alpha_{jg,y_{jg}+1} - \theta_{jg}z_0 - \delta_{jg}\sqrt{1-\theta_{jg}^2}z_g}{\sqrt{(1-\theta_{jg}^2)(1-\delta_{jg}^2)}}\right)\right.\right. \\
&\quad \left.\left.-\Phi\left(\frac{\alpha_{jg,y_{jg}} - \theta_{jg}z_0 - \delta_{jg}\sqrt{1-\theta_{jg}^2}z_g}{\sqrt{(1-\theta_{jg}^2)(1-\delta_{jg}^2)}}\right)\right] \phi(z_g)dz_g\right\} \phi(z_0)dz_0.
\end{aligned}$$

This model is the same as the bi-factor Gaussian model (Gibbons and Hedeker, 1992; Gibbons et al., 2007) with stochastic representation

$$Z_{jg} = \theta_{jg}Z_0 + \gamma_{jg}Z_g + \sqrt{1-\theta_{jg}^2-\gamma_{jg}^2}\epsilon_{jg}, \qquad g = 1,\ldots,G, \quad j = 1,\cdots,d_g, \tag{3.3}$$

where $\gamma_{jg} = \delta_{jg}\sqrt{1-\theta_{jg}^2}$ and Z_0, Z_g, ϵ_{jg} are i.i.d. $N(0,1)$ random variables. The parameter θ_{jg} of C_{Y_{jg},X_0} is the correlation of Z_{jg} and Z_0, and the parameter δ_{jg} of $C_{Y_{jg},X_g|X_0}$ is the partial correlation between Z_{jg} and $Z_g = \Phi^{-1}(X_g)$ given $Z_0 = \Phi^{-1}(X_0)$.

It implies that the underlying random variables Z_{jg}'s have a multivariate Gaussian distribution where the off-diagonal entries of the correlation matrix have the

form $\theta_{j_1 g}\theta_{j_2 g} + \gamma_{j_1 g}\gamma_{j_2 g}$ and $\theta_{j_1 g_1}\theta_{j_2 g_2}$ for $j_1 \neq j_2$ and $g_1 \neq g_2$, respectively. For the Gaussian bi-factor model to be identifiable, the number of dependence parameters has to be $2d - N_1 - N_2$, where N_1 and N_2 is the number of groups that consist of 1 and 2 items, respectively. For a group g of size 1 with variable j, Z_g is absorbed with ϵ_{jg} because γ_{jg} would not be identifiable. For a group g of size 2 with variable indices j_1, j_2, the parameters $\gamma_{j_1 g}$ and $\gamma_{j_2 g}$ appear only in one correlation, hence one of $\gamma_{j_1 g}, \gamma_{j_2 g}$ can be taken as 1 without loss of generality. For the bi-factor copula with non-Gaussian linking copulas, near non-identifiability can occur when there are groups of size 2; in this case, one of the linking copulas to the group latent variable can be fixed at comonotonicity.

For the Gaussian second-order model let $Z_0, Z_1', \ldots, Z_G'$ be the dependent latent $N(0,1)$ variables, where Z_0 is the second-order factor and $Z_g' = \beta_g Z_0 + \sqrt{1-\beta_g^2} Z_g$ is the first-order factor for group g. That is, there is an one second-order factor Z_0, and the first-order factors $Z_1', \ldots, Z_G'$ are linear combinations of the second-order factor, plus a unique variable Z_g for each first-order factor. The stochastic representation is (Krupskii and Joe, 2015):

$$\begin{aligned} Z_{jg} &= \beta_{jg} Z_g' + \sqrt{1-\beta_{jg}^2}\epsilon_{jg} \\ Z_g' &= \beta_g Z_0 + \sqrt{1-\beta_g^2} Z_g, \qquad g = 1, \ldots, G, \quad j = 1, \cdots, d_g, \end{aligned}$$

or

$$Z_{jg} = \beta_{jg}\beta_g Z_0 + \beta_{jg}\sqrt{1-\beta_g^2} Z_g + \sqrt{1-\beta_{jg}^2}\epsilon_{jg}, \quad j = 1, \cdots, d_g. \tag{3.4}$$

Hence, this is a special case of (3.3) where $\theta_{jg} = \beta_{jg}\beta_g$ and $\gamma_{jg} = \beta_{jg}\sqrt{1-\beta_g^2}$.

3.2 Estimation and computational details

For the set of all parameters, let $\boldsymbol{\theta} = (\mathbf{a}, \boldsymbol{\theta}_g, \boldsymbol{\delta}_g)$ for the bi-factor copula model and $\boldsymbol{\theta} = (\mathbf{a}, \boldsymbol{\theta}_g, \boldsymbol{\delta})$ for the second-order copula model, where $\mathbf{a} = (a_{jg,k} : j = 1, \ldots, d_g, g = 1, \ldots, G, k = 1, \ldots, K-1)$, $\boldsymbol{\theta}_g = (\theta_{1g}, \ldots, \theta_{jg}, \ldots, \theta_{d_g g} : g = 1, \ldots, G)$, $\boldsymbol{\delta}_g = (\delta_{1g}, \ldots, \delta_{jg}, \ldots, \delta_{d_g g} : g = 1, \ldots, G)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_G)$.

With sample size n and data $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the joint log-likelihood of the bi-factor and second-order copula is

$$\ell(\boldsymbol{\theta}; \mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^{n} \log \pi(\mathbf{y}_i; \boldsymbol{\theta}), \tag{3.5}$$

with $\pi(\mathbf{y}_i; \boldsymbol{\theta})$ as in (3.1) and (3.2), respectively. Maximization of (3.5) is numerically possible but time-consuming for large d because of many univariate cutpoints and dependence parameters. Hence, we approach estimation using the two-step IFM method proposed by Joe (2005) that can efficiently (in the sense of computing time and asymptotic variance) estimate the model parameters.

In the first step, the cutpoints are estimated using the univariate sample proportions. The univariate cutpoints for the jth item in group g are estimated as $\hat{a}_{jg,k} = \sum_{y=0}^{k} p_{jg,y}$, where $p_{jg,y}, y = 0, \ldots, K-1$ for $g = 1, \ldots, G$ and $j = 1, \ldots, d_g$ are the univariate sample proportions. In the second step of the IFM method, the joint log-likelihood in (3.5) is maximized over the copula parameters with the cutpoints fixed as estimated at the first step. The estimated copula parameters can be obtained

by using a quasi-Newton (Nash, 1990) method applied to the logarithm of the joint likelihood.

For the bi-factor copula model numerical evaluation of the joint pmf can be achieved with the following steps:

1. Calculate Gauss-Legendre quadrature (Stroud and Secrest, 1966) points $\{x_q : q = 1, \ldots, n_q\}$ and weights $\{w_q : q = 1, \ldots, n_q\}$ in terms of standard uniform.

2. Numerically evaluate the joint pmf

$$\int_0^1 \prod_{g=1}^{G} \left\{ \int_0^1 \prod_{j=1}^{d_g} f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g, x_0) dx_g \right\} dx_0$$

in a double sum

$$\sum_{q_1=1}^{n_q} w_{q_1} \prod_{g=1}^{G} \left\{ \sum_{q_2=1}^{n_q} w_{q_2} \prod_{j=1}^{d_g} f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_{q_2}, x_{q_1}) \right\}$$

For the second-order copula model numerical evaluation of the joint pmf can be achieved with the following steps:

1. Calculate Gauss-Legendre quadrature points $\{x_q : q = 1, \ldots, n_q\}$ and weights $\{w_q : q = 1, \ldots, n_q\}$ in terms of stand uniform.

2. Numerically evaluate the joint pmf

$$\int_0^1 \left\{ \prod_{g=1}^{G} \int_0^1 \Big[\prod_{j=1}^{d_g} f_{Y_{jg}|X_g}(y_{jg}|x_g; \theta_{jg}) \Big] c_{X_g, X_0}\big(x_g, x_0; \delta_g\big) dx_g \right\} dx_0$$

in a double sum

$$\sum_{q_1=1}^{n_q} w_{q_1} \left\{ \prod_{g=1}^{G} \sum_{q_2=1}^{n_q} w_{q_2} \Big[\prod_{j=1}^{d_g} f_{Y_{jg}|X_g}(y_{jg}|x_{q_2|q_1};\theta_{jg}) \Big] \right\}$$

where $x_{q_2|q_1} = C^{-1}_{Y_{jg}|X_g;X_0}(x_{q_2}|x_{q_1};\delta_g)$. Note that the independent quadrature points $\{x_{q_1} : q_1 = 1,\ldots,n_q\}$ and $\{x_{q_2} : q_2 = 1,\ldots,n_q\}$ have been converted to dependent quadrature points that have an one-factor copula distribution $C_X(\cdot;\boldsymbol{\delta})$.

Our comparisons show that $n_q = 25$ quadrature points provide good precision for both the bi-factor and second-order copula models.

3.3 Bivariate copula selection

In line with Nikoloulopoulos and Joe (2015) and as discussed in the Introduction 1.5.2, we use bivariate parametric copulas that can be used when considering latent maxima, minima or mixtures of means, namely the Gumbel, survival Gumbel (s.Gumbel) and Student t_ν copulas, respectively. A model with bivariate Gumbel copulas that possess upper tail dependence has latent (ordinal) variables that can be considered as (discretized) maxima, and there is more probability in the joint upper tail. A model with bivariate s.Gumbel copulas that possess lower tail dependence has latent (ordinal) variables that can be considered as (discretized) minima, and there is more probability in the joint lower tail. A model with bivariate t_ν copulas that possess the same lower and upper tail dependence has latent (ordinal) variables that can be considered as mixtures of (discretized) means, since the bivariate Student t_ν

distribution arises as a scale mixture of bivariate normals. A small value of ν, such as $1 \leq \nu \leq 5$, leads to a model with more probabilities in the joint upper and joint lower tails compared with the BVN copula.

In the following subsection we describe a heuristic method that automatically selects the bivariate parametric copula families that build either the bi-factor or the second-order copula model. In the context of items that can be split into G non-overlapping groups, such that there is homogeneous dependence within each group, it is sufficient to (a) summarize the average of the polychoric semi-correlations presented in the previous chapter in Section 2.1.1 for all pairs within each of the G groups and for all pairs of items, and (b) not mix bivariate copulas for a single factor; hence, for both the bi-factor and second-order copula models we allow $G + 1$ different copula families, one for each group specific factor X_g and one for X_0.

3.3.1 Selection algorithm

We propose a heuristic method that selects appropriate bivariate copulas for each factor of the bi-factor and second-order copula models. It starts with an initial assumption, that all bivariate linking copulas are BVN copulas, i.e. the starting model is either the Gaussian bi-factor or second-order model, and then sequentially other copulas with lower or upper tail dependence are assigned to the factors where necessary to account for more probability in one or both joint tails. The selection algorithm involves the following steps:

1. Fit the bi-factor or second-order copula model with BVN copulas.

2. Fit all the possible bi-factor or second-order copula models, iterating over all the copula candidates that link all items Y_{jg}'s in group g or each group-specific factor X_g, respectively, to X_0.

3. Select the copula family that corresponds to the lowest Akaike information criterion (AIC), that is, AIC $= -2 \times \ell + 2 \times \#$copula parameters.

4. Fix the selected copula family that links the observed (bi-factor model) or latent (second-order model) variables to X_0.

5. For $g = 1, \ldots, G$:

   (a) Fit all the possible models, iterating over all the copula candidates that link all the items in group g to the group-specific factor X_g.

   (b) Select the copula family that corresponds to the lowest AIC.

   (c) Fix the selected linking copula family for all the items in group g with X_g.

3.4 Goodness-of-fit

We will use the limited information M_2 statistic proposed by Maydeu-Olivares and Joe (2006) to evaluate the overall fit of the proposed bi-factor and second-order copula models. The M_2 statistic is based on a quadratic form of the deviations of sample and model-based proportions over all bivariate margins. It has been utilised for factor copula models for item response data (Nikoloulopoulos and Joe, 2015), and for

mixed continuous and discrete data (Kadhem and Nikoloulopoulos, 2021b) as shown in Chapter 2. The M_2 statistic has been summarised and discussed in Section 2.4.2.

The M_2 involves the first order derivatives of the univariate and bivariate marginal probabilities with respect to the estimated model parameters. We summarise the form of the derivatives of the univariate and bivariate marginal probabilities with respect to the estimated model parameters in below tables for the bi-factor and second-order copula models. Table 3.1 gives the derivatives of the univariate probability with respect to the cutpoints. The derivatives of the bivariate margins with respect to the cutpoints and copula parameters for the bi-factor copula model are given in Table 3.2 if bivariate margins are in the same group, and in Table 3.3 if bivariate margins are in different non-overlapping groups. The derivatives of the bivariate margins with respect to the cutpoints and copula parameters for the second-order copula model are given in Table 3.4 if bivariate margins are within the same group, and in Table 3.5 if bivariate margins are in different non-overlapping groups.

Table 3.1: Derivatives of the univariate probability $\pi_{jg,y} = \Phi(\alpha_{jg,y+1}) - \Phi(\alpha_{jg,y})$ with respect to the cutpoint $\alpha_{jg,k}$ for $g = 1 \ldots, G,\ j = 1, \ldots, d_g,\ y = 1, \ldots, K-1,$ and $k = 1, \ldots, K-1$.

$\partial \pi_{jg,y} / \partial \alpha_{jg,k}$	If
$\phi(\alpha_{jg,y+1})$	$k = y+1$
$-\phi(\alpha_{jg,y})$	$k = y$

Table 3.2: Derivatives of the bivariate probability $\pi_{j_1 j_2 g, y_1, y_2} = \Pr(Y_{j_1 g} = y_1, Y_{j_2 g} = y_2)$ with respect to the cutpoint $\alpha_{jg,k}$, the copula parameter θ_{jg} for the common factor X_0, and the copula parameter δ_{jg} for the group-specific factor X_g for the bi-factor copula model for $g = 1 \ldots, G,\ j, j_1, j_2 = 1, \ldots, d_g,\ y, y_1, y_2 = 1, \ldots, K-1,$ and $k = 1, \ldots, K-1$.

$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \alpha_{jg,k}$	If
$\phi(\alpha_{j_1 g, y_1+1}) \int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g;X_0}(y_{j_2 g}\|x_g;x_0)\ c_{X_g Y_{j_1 g}}\big(x_g, C_{Y_{j_1 g}\|X_0}(a_{j_1 g, y_1+1}\|x_0)\big)\ c_{X_0 Y_{j_1 g}}(x_0, a_{j_1 g, y_1+1})\ dx_g dx_0$	$j = j_1, k = y_1 + 1$
$-\phi(\alpha_{j_1 g, y_1}) \int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g;X_0}(y_{j_2 g}\|x_g;x_0)\ c_{X_g Y_{j_1 g}}\big(x_g, C_{Y_{j_1 g}\|X_0}(a_{j_1 g, y_1}\|x_0)\big)\ c_{X_0 Y_{j_1 g}}(x_0, a_{j_1 g, y_1})\ dx_g dx_0$	$j = j_1, k = y_1$
$\phi(\alpha_{j_2 g, y_2+1}) \int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g;X_0}(y_{j_1 g}\|x_g;x_0)\ c_{X_g Y_{j_2 g}}\big(x_g, C_{Y_{j_2 g}\|X_0}(a_{j_2 g, y_2+1}\|x_0)\big)\ c_{X_0 Y_{j_2 g}}(x_0, a_{j_2 g, y_2+1})\ dx_g dx_0$	$j = j_2, k = y_2 + 1$
$-\phi(\alpha_{j_2 g, y_2}) \int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g;X_0}(y_{j_1 g}\|x_g;x_0)\ c_{X_g Y_{j_2 g}}\big(x_g, C_{Y_{j_2 g}\|X_0}(a_{j_2 g, y_2}\|x_0)\big)\ c_{X_0 Y_{j_2 g}}(x_0, a_{j_2 g, y_2})\ dx_g dx_0$	$j = j_2, k = y_2$
$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \theta_{jg}$	If
$\int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g;X_0}(y_{j_2 g}\|x_g;x_0)\ \bar{f}_{Y_{j_1 g}\|X_{j_1 g};X_0}(y_{j_1 g}\|x_g;x_0)\ dx_g dx_0$	$j = j_1$
$\int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g;X_0}(y_{j_1 g}\|x_g;x_0)\ \bar{f}_{Y_{j_2 g}\|X_{j_2 g};X_0}(y_{j_2 g}\|x_g;x_0)\ dx_g dx_0$	$j = j_2$
$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \delta_{jg}$	If
$\int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g;X_0}(y_{j_2 g}\|x_g;x_0)\ \dot{f}_{Y_{j_1 g}\|X_{j_1 g};X_0}(y_{j_1 g}\|x_g;x_0)\ dx_g dx_0$	$j = j_1$
$\int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g;X_0}(y_{j_1 g}\|x_g;x_0)\ \dot{f}_{Y_{j_2 g}\|X_{j_2 g};X_0}(y_{j_2 g}\|x_g;x_0)\ dx_g dx_0$	$j = j_2$

Note that $f_{Y_{jg}|X_g;X_0}(y_{jg}|x_g;x_0) = \Big(C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0;\theta_{jg})|x_g;\delta_{jg}\big) - C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y}|x_0;\theta_{jg})|x_g;\delta_{jg}\big)\Big)$ where $a_{jg,k} = \Phi(\alpha_{jg,k})$, $c_{X_0 Y_{jg}}(x_0, a) = \partial^2 C_{X_0 Y_{jg}}(x_0, a)/\partial x_0 \partial a$, $\dot{C}_{jg|X_0}(\cdot;\theta_{jg}) = \partial C_{jg|X_0}(\cdot;\theta_{jg})/\partial \theta_{jg}$, $\dot{C}_{Y_{jg}|X_g;X_0}(\cdot;\delta_{jg}) = \partial C_{Y_{jg}|X_g;X_0}(\cdot;\delta_{jg})/\partial \delta_{jg}$, $\dot{f}_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0) = \partial f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0)/\partial \delta_{jg} = \dot{C}_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0)|x_g\big) - \dot{C}_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y}|x_0)|x_g\big)$, $\bar{f}_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0) = \partial f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0)/\partial \theta_{jg} = c_{X_g Y_{jg}}\big(x_g, C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0)\big)\dot{C}_{Y_{jg}|X_0}(a_{jg,y+1}|x_0) - c_{X_g Y_{jg}}\big(x_g, C_{Y_{jg}|X_0}(a_{jg,y}|x_0)\big)\dot{C}_{Y_{jg}|X_0}(a_{jg,y}|x_0)$.

Table 3.3: Derivatives of the bivariate probability $\pi_{j_1g_1j_2g_2,y_1,y_2} = \Pr(Y_{j_1g_1} = y_1, Y_{j_2g_2} = y_2)$ with respect to the cutpoint $\alpha_{jg,k}$, the copula parameter θ_{jg} for the common factor X_0, and the copula parameter δ_{jg} for the group-specific factor X_g for the bi-factor copula model for $g = 1\ldots, G,\ j, j_1, j_2 = 1, \ldots, d_g,\ y, y_1, y_2 = 1, \ldots, K-1,$ and $k = 1, \ldots, K-1$.

$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\alpha_{jg,k}$	If
$\phi(\alpha_{j_1g_1,y_1+1})\int_0^1\int_0^1 f_{Y_{j_2g_2}\|X_{g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)dx_{g_2}\int_0^1 c_{X_{g_1}Y_{j_1g_1}}\big(x_{g_1}, C_{Y_{j_1g_1}\|X_0}(a_{j_1g_1,y_1+1}\|x_0)\big)c_{X_0Y_{j_1g_1}}(x_0, a_{j_1g_1,y_1+1})dx_{g_1}dx_0$	$j = j_1, g = g_1, k = y_1+1$
$-\phi(\alpha_{j_1g_1,y_1})\int_0^1\int_0^1 f_{Y_{j_2g_2}\|X_{g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)dx_{g_2}\int_0^1 c_{X_{g_1}Y_{j_1g_1}}\big(x_{g_1}, C_{Y_{j_1g_1}\|X_0}(a_{j_1g_1,y_1}\|x_0)\big)c_{X_0Y_{j_1g_1}}(x_0, a_{j_1g_1,y_1})dx_{g_1}dx_0$	$j = j_1, g = g_1, k = y_1$
$\phi(\alpha_{j_2g_2,y_2+1})\int_0^1\int_0^1 f_{Y_{j_1g_1}\|X_{g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)dx_{g_1}\int_0^1 c_{X_{g_2}Y_{j_2g_2}}\big(x_{g_2}, C_{Y_{j_2g_2}\|X_0}(a_{j_2g_2,y_2+1}\|x_0)\big)c_{X_0Y_{j_2g_2}}(x_0, a_{j_2g_2,y_2+1})dx_{g_2}dx_0$	$j = j_2, g = g_2, k = y_2+1$
$-\phi(\alpha_{j_2g_2,y_2})\int_0^1\int_0^1 f_{Y_{j_1g_1}\|X_{g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)dx_{g_1}\int_0^1 c_{X_{g_2}Y_{j_2g_2}}\big(x_{g_2}, C_{Y_{j_2g_2}\|X_0}(a_{j_2g_2,y_2}\|x_0)\big)c_{X_0Y_{j_2g_2}}(x_0, a_{j_2g_2,y_2})dx_{g_2}dx_0$	$j = j_2, g = g_2, k = y_2$
$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\theta_{jg}$	If
$\int_0^1\int_0^1 f_{Y_{j_2g_2}\|X_{g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)\ dx_{g_2}\int_0^1 \bar{f}_{Y_{j_1g_1}\|X_{j_1g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)\ dx_{g_1}dx_0$	$j = j_1, g = g_1$
$\int_0^1\int_0^1 f_{Y_{j_1g_1}\|X_{g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)\ dx_{g_1}\int_0^1 \bar{f}_{Y_{j_2g_2}\|X_{j_2g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)\ dx_{g_2}dx_0$	$j = j_2, g = g_2$
$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\delta_{jg}$	If
$\int_0^1\int_0^1 f_{Y_{j_2g_2}\|X_{g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)\ dx_{g_2}\int_0^1 \dot{f}_{Y_{j_1g_1}\|X_{j_1g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)\ dx_{g_1}dx_0$	$j = j_1, g = g_1$
$\int_0^1\int_0^1 f_{Y_{j_1g_1}\|X_{g_1};X_0}(y_{j_1g_1}\|x_{g_1};x_0)\ dx_{g_1}\int_0^1 \dot{f}_{Y_{j_2g_2}\|X_{j_2g_2};X_0}(y_{j_2g_2}\|x_{g_2};x_0)\ dx_{g_2}dx_0$	$j = j_2, g = g_2$

Note that $f_{Y_{jg}|X_g;X_0}(y_{jg}|x_g;x_0) = \Big(C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0;\theta_{jg})|x_g;\delta_{jg}\big) - C_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y}|x_0;\theta_{jg})|x_g;\delta_{jg}\big)\Big)$ where $a_{jg,k} = \Phi(\alpha_{jg,k})$, $c_{X_0Y_{jg}}(x_0,a) = \partial^2 C_{X_0Y_{jg}}(x_0,a)/\partial x_0\partial a$, $\dot{C}_{jg|X_0}(\cdot;\theta_{jg}) = \partial C_{jg|X_0}(\cdot;\theta_{jg})/\partial\theta_{jg}$, $\dot{C}_{Y_{jg}|X_g;X_0}(\cdot;\delta_{jg}) = \partial C_{Y_{jg}|X_g;X_0}(\cdot;\delta_{jg})/\partial\delta_{jg}$, $\dot{f}_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0) = \partial f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0)/\partial\delta_{jg} = \dot{C}_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0)|x_g\big) - \dot{C}_{Y_{jg}|X_g;X_0}\big(C_{Y_{jg}|X_0}(a_{jg,y}|x_0)|x_g\big)$, $\bar{f}_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0) = \partial f_{Y_{jg}|X_{jg};X_0}(y_{jg}|x_g;x_0)/\partial\theta_{jg} = c_{X_gY_{jg}}\big(x_g, C_{Y_{jg}|X_0}(a_{jg,y+1}|x_0)\big)\dot{C}_{Y_{jg}|X_0}(a_{jg,y+1}|x_0) - c_{X_gY_{jg}}\big(x_g, C_{Y_{jg}|X_0}(a_{jg,y}|x_0)\big)\dot{C}_{Y_{jg}|X_0}(a_{jg,y}|x_0)$.

Table 3.4: Derivatives of the bivariate probabilities $\pi_{j_1 j_2 g, y_1, y_2} = \Pr(Y_{j_1 g} = y_1, Y_{j_2 g} = y_2)$ with respect to the cutpoint $\alpha_{jg,k}$, the copula parameter θ_{jg} for the first-order factor X_g, and the copula parameter δ_g for the the second-order factor X_0 for the second-order copula model for $g = 1 \ldots, G,\ j, j_1, j_2 = 1, \ldots, d_g,\ y, y_1, y_2 = 1, \ldots, K-1,$ and $k = 1, \ldots, K-1$.

$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \alpha_{jg,k}$	If
$\phi(\alpha_{j_1 g, y_1+1}) \int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g}(y_{j_2 g}\|x_g)\ c_{X_g Y_{j_1 g}}(x_g, a_{j_1 g, y_1+1})\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_1, k = y_1 + 1$
$-\phi(\alpha_{j_1 g, y_1})\ \int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g}(y_{j_2 g}\|x_g)\ c_{X_g Y_{j_1 g}}(x_g, a_{j_1 g, y_1})\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_1, k = y_1$
$\phi(\alpha_{j_2 g, y_2+1}) \int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g}(y_{j_1 g}\|x_g)\ c_{X_g Y_{j_2 g}}(x_g, a_{j_2 g, y_2+1})\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_2, k = y_2 + 1$
$-\phi(\alpha_{j_2 g, y_2})\ \int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g}(y_{j_1 g}\|x_g)\ c_{X_g Y_{j_2 g}}(x_g, a_{j_2 g, y_2})\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_2, k = y_2$
$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \theta_{jg}$	If
$\int_0^1 \int_0^1 \ f_{Y_{j_2 g}\|X_g}(y_{j_2 g}\|x_g)\ \dot{f}_{Y_{j_1 g}\|X_g}(y_{j_1 g}\|x_g)\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_1$
$\int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g}(y_{j_1 g}\|x_g)\ \dot{f}_{Y_{j_2 g}\|X_g}(y_{j_2 g}\|x_g)\ c_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	$j = j_2$
$\partial \pi_{j_1 j_2 g, y_1, y_2} / \partial \delta_g$	
$\int_0^1 \int_0^1 \ f_{Y_{j_1 g}\|X_g}(y_{j_1 g}\|x_g)\ f_{Y_{j_2 g}\|X_g}(y_{j_2 g}\|x_g)\ \dot{c}_{X_g X_0}(x_g, x_0)\ dx_g dx_0$	

Note that $f_{Y_{jg}|X_g}(y_{jg}|x_g) = C_{Y_{jg}|X_g}\big(a_{jg,y+1}|x_g;\theta_{jg}\big) - C_{Y_{jg}|X_g}\big(a_{jg,y}|x_g;\theta_{jg}\big)$, $c_{X_g Y_{jg}}(x_g, a) = \partial^2 C_{X_g Y_{jg}}(x_g, a)/\partial x_g \partial a$, $\dot{C}_{Y_{jg}|X_g}(\cdot;\theta_{jg}) = \partial C_{Y_{jg}|X_g}(\cdot;\theta_{jg})/\partial\theta_{jg}$, $\dot{f}_{Y_{jg}|X_{jg}}(y_{jg}|x_g) = \partial f_{Y_{jg}|X_{jg}}(y_{jg}|x_g)/\partial\theta_{jg} = \dot{C}_{Y_{jg}|X_g}\big(a_{jg,y+1}|x_g\big) - \dot{C}_{Y_{jg}|X_g}\big(a_{jg,y}|x_g\big)$, $\dot{c}_{X_g X_0}(x_g, x_0; \delta_g) = \partial c_{X_g X_0}(x_g, x_0; \delta_g)/\partial\delta_g$.

Table 3.5: Derivatives of the bivariate probability $\pi_{j_1g_1j_2g_2,y_1,y_2} = \Pr(Y_{j_1g_1} = y_1, Y_{j_2g_2} = y_2)$ with respect to the cutpoint $\alpha_{jg,k}$, the copula parameter θ_{jg} for the first-order factor X_g, and the copula parameter δ_g for the second-order factor X_0 for the second-order copula model for $g = 1\ldots, G,\ j, j_1, j_2 = 1, \ldots, d_g,\ y, y_1, y_2 = 1, \ldots, K-1,$ and $k = 1, \ldots, K-1$.

$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\alpha_{jg,k}$	If
$\phi(\alpha_{j_1g_1,y_1+1}) \int_0^1 \int_0^1 \ f_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2} \int_0^1 \ c_{X_{g_1}Y_{j_1g_1}}(x_{g_1},a_{j_1g_1,y_1+1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1}dx_0$	$j = j_1, g = g_1, k = y_1 + 1$
$-\phi(\alpha_{j_1g_1,y_1}) \int_0^1 \int_0^1 \ f_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2} \int_0^1 \ c_{X_{g_1}Y_{j_1g_1}}(x_{g_1},a_{j_1g_1,y_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1}dx_0$	$j = j_1, g = g_1, k = y_1$
$\phi(\alpha_{j_2g_2,y_2+1}) \int_0^1 \int_0^1 \ f_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1} \int_0^1 \ c_{X_{g_2}Y_{j_2g_2}}(x_{g_2},a_{j_2g_2,y_2+1})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2}dx_0$	$j = j_2, g = g_2, k = y_2 + 1$
$-\phi(\alpha_{j_2g_2,y_2}) \int_0^1 \int_0^1 \ f_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1} \int_0^1 \ c_{X_{g_2}Y_{j_2g_2}}(x_{g_2},a_{j_2g_2,y_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2}dx_0$	$j = j_2, g = g_2, k = y_2$
$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\theta_{jg}$	If
$\int_0^1 \int_0^1 \ f_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2} \int_0^1 \ \dot{f}_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1}dx_0$	$j = j_1, g = g_1$
$\int_0^1 \int_0^1 \ f_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1} \int_0^1 \ \dot{f}_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2}dx_0$	$j = j_2, g = g_2$
$\partial\pi_{j_1g_1j_2g_2,y_1,y_2}/\partial\delta_g$	If
$\int_0^1 \int_0^1 \ f_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ c_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2} \int_0^1 \ f_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ \dot{c}_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1}dx_0$	$g = g_1$
$\int_0^1 \int_0^1 \ f_{Y_{j_1g_1}\|X_{g_1}}(y_{j_1g_1}\|x_{g_1})\ c_{X_{g_1}X_0}(x_{g_1},x_0)\ dx_{g_1} \int_0^1 \ f_{Y_{j_2g_2}\|X_{g_2}}(y_{j_2g_2}\|x_{g_2})\ \dot{c}_{X_{g_2}X_0}(x_{g_2},x_0)\ dx_{g_2}dx_0$	$g = g_2$

Note that $f_{Y_{jg}|X_g}(y_{jg}|x_g) = C_{Y_{jg}|X_g}\big(a_{jg,y+1}|x_g;\theta_{jg}\big) - C_{Y_{jg}|X_g}\big(a_{jg,y}|x_g;\theta_{jg}\big)$, $c_{X_gY_{jg}}(x_g,a) = \partial^2 C_{X_gY_{jg}}(x_g,a)/\partial x_g\partial a$, $\dot{C}_{Y_{jg}|X_g}(\cdot;\theta_{jg}) = \partial C_{Y_{jg}|X_g}(\cdot;\theta_{jg})/\partial\theta_{jg}$, $\dot{f}_{Y_{jg}|X_{jg}}(y_{jg}|x_g) = \partial f_{Y_{jg}|X_{jg}}(y_{jg}|x_g)/\partial\theta_{jg} = \dot{C}_{Y_{jg}|X_g}\big(a_{jg,y+1}|x_g\big) - \dot{C}_{Y_{jg}|X_g}\big(a_{jg,y}|x_g\big)$, $\dot{c}_{X_gX_0}(x_g,x_0;\delta_g) = \partial c_{X_gX_0}(x_g,x_0;\delta_g)/\partial\delta_g$.

3.5 Simulations

An extensive simulation study is conducted to (a) gauge the small-sample efficiency of the IFM estimation method and investigate the misspecification of the bivariate pair-copulas, (b) examine the reliability of using the heuristic algorithm to select the true (simulated) bivariate linking copulas, and (c) study the small-sample performance of the M_2 statistic.

We randomly generate 1,000 datasets with samples of size $n = 500$ or 1000 and $d = 16$ items, with $K = 3$ or $K = 5$ equally weighted categories, that are equally separated into $G = 4$ non-overlapping groups from the bi-factor and second-order copula model. In each simulated model, we use different linking copulas to cover different types of dependence. To make the models comparable, we convert the BVN/t_ν and Gumbel/s.Gumbel copula parameters to Kendall's τ's via

$$\tau(\theta) = \frac{2}{\pi} \arcsin(\theta) \tag{3.6}$$

and

$$\tau(\theta) = 1 - \theta^{-1}, \tag{3.7}$$

respectively. For the bi-factor copula models we set $\tau(\boldsymbol{\theta}_g) = (0.45, 0.55, 0.65, 0.75)$ and $\tau(\boldsymbol{\delta}_g) = (0.30, 0.35,\ 0.40, 0.50)$ for $g = 1, \ldots, 4$. For the second-order copula models we set $\tau(\boldsymbol{\theta}_g) = (0.4, 0.5, 0.6, 0.7)$ for $g = 1, \ldots, 4$ and $\tau(\boldsymbol{\delta}) = (0.30, 0.35, 0.40, 0.45)$.

The Kendall's tau parameters $\tau(\boldsymbol{\theta}_g)$ and $\tau(\boldsymbol{\delta}_g)$ as described above are common for each group, hence Table 3.6 contains the group estimated average biases, root mean square errors (RMSE), and standard deviations (SD), scaled by n, for the IFM

Table 3.6: Small sample of size $n = 500$ simulations (10^3 replications) from the bi-factor and second-order factor models with Gumbel copulas and group estimated average biases, root mean square errors (RMSE), and standard deviations (SD), scaled by n, for the IFM estimates under different pair-copulas from the bi-factor and second-order copula models.

			Bi-factor copula model								Second-order copula model							
			$\tau(\boldsymbol{\theta}_g),\ g = 1,\ldots,4$				$\tau(\boldsymbol{\delta}_g),\ g = 1,\ldots,4$				$\tau(\boldsymbol{\delta})$				$\tau(\boldsymbol{\theta}_g),\ g = 1,\ldots,4$			
	Fitted model	K	0.45	0.55	0.65	0.75	0.30	0.35	0.40	0.50	0.30	0.35	0.40	0.45	0.40	0.50	0.60	0.70
$n$bias	BVN	3	2.65	2.54	2.66	2.16	6.60	7.81	6.99	6.39	5.58	5.34	5.33	5.60	0.41	0.86	0.62	0.27
		5	1.98	2.27	2.54	2.53	5.99	6.27	5.42	2.31	8.71	8.36	7.94	8.52	0.93	0.51	0.58	2.52
	Gumbel	3	0.39	0.35	0.28	0.34	0.89	1.02	1.62	3.40	-0.18	0.18	0.18	1.88	0.22	0.67	1.14	2.37
		5	0.23	0.23	0.07	0.20	0.84	0.85	1.02	1.98	0.22	0.13	-0.25	1.15	0.23	0.43	0.63	0.60
	s.Gumbel	3	3.59	3.03	1.51	0.31	4.86	4.52	4.21	1.19	18.43	18.29	18.54	18.68	6.32	6.18	5.47	3.67
		5	0.79	2.25	3.80	5.30	15.89	15.82	13.89	14.52	25.65	24.80	23.58	22.59	3.77	2.54	1.24	2.74
	t_5	3	1.65	2.81	3.28	3.48	6.99	8.20	7.07	4.89	7.98	8.55	9.18	9.55	3.36	3.56	4.71	3.81
		5	0.49	0.49	0.84	0.92	5.81	6.09	5.58	1.69	9.71	10.05	9.82	9.87	2.24	2.29	2.64	0.36
nSE	BVN	3	15.03	13.42	12.37	11.06	30.77	31.20	33.07	39.93	22.80	24.94	24.97	27.03	16.82	16.41	17.06	21.32
		5	13.68	11.89	10.63	8.95	24.58	25.33	25.70	29.86	21.28	23.04	22.45	24.72	15.09	14.27	14.01	15.41
	Gumbel	3	15.10	13.81	12.33	10.97	29.61	31.34	32.82	42.17	22.58	24.73	25.35	27.87	16.99	16.73	17.66	22.02
		5	13.67	12.29	10.55	8.76	23.60	24.72	25.39	31.13	20.75	22.75	22.69	24.86	15.31	14.62	14.33	15.72
	s.Gumbel	3	15.58	13.76	12.60	11.27	33.77	34.80	38.18	51.31	25.34	26.80	27.19	29.36	17.40	16.49	16.59	18.46
		5	14.11	12.30	11.16	9.66	27.08	28.44	30.18	40.10	22.61	24.13	23.36	25.46	15.90	14.57	14.38	16.89
	t_5	3	15.29	13.54	12.27	10.79	31.43	31.74	33.02	39.02	23.59	25.57	25.65	27.61	17.48	16.69	17.64	22.03
		5	13.84	11.99	10.55	8.80	24.79	25.35	25.66	29.10	21.67	23.52	22.67	24.52	15.40	14.52	14.03	14.88
nRMSE	BVN	3	15.28	13.66	12.66	11.27	31.48	32.19	33.81	40.45	23.47	25.50	25.53	27.60	16.83	16.44	17.08	21.33
		5	13.83	12.11	10.93	9.30	25.31	26.10	26.27	29.96	22.99	24.51	23.81	26.14	15.12	14.28	14.03	15.62
	Gumbel	3	15.10	13.81	12.34	10.98	29.63	31.37	32.87	42.31	22.58	24.73	25.35	27.94	16.99	16.75	17.71	22.15
		5	13.67	12.30	10.55	8.77	23.62	24.74	25.42	31.20	20.75	22.75	22.69	24.88	15.31	14.63	14.35	15.73
	s.Gumbel	3	16.00	14.09	12.69	11.27	34.13	35.13	38.42	51.33	31.33	32.45	32.91	34.80	18.52	17.65	17.49	18.82
		5	14.14	12.51	11.79	11.02	31.41	32.55	33.22	42.67	34.19	34.60	33.19	34.04	16.35	14.82	14.44	17.13
	t_5	3	15.40	13.83	12.71	11.34	32.21	32.80	33.77	39.32	24.91	26.97	27.24	29.21	17.80	17.08	18.27	22.36
		5	13.85	12.01	10.59	8.86	25.47	26.08	26.26	29.16	23.75	25.58	24.71	26.43	15.56	14.71	14.29	14.89

estimates under different pair-copulas from the bi-factor and second-order copula models. In the true (simulated) models the linking copulas are Gumbel copulas.

Conclusions from the values in the table are the following:

- IFM with the true bi-factor or second-order model is highly efficient according to the simulated biases, SDs and RMSEs.
- The IFM estimates of τ's are not robust under copula misspecification and their biases increase when the assumed bivariate copula has tail dependence of opposite direction from the true bivariate copula. For example, in Table 3.6 the scaled biases for the IFM estimates increase substantially when the linking copulas are the s.Gumbel copulas.

To examine the reliability of using the heuristic algorithm to select the true (simulated) bivariate linking copulas, samples of size 500 were generated from various bi-factor and second-order copula models. Table 3.7 presents the number of times that the true (simulated) linking copulas were chosen over 1,000 simulation runs. It is revealed that the model selection algorithm performs extremely well for various bi-factor and second-order copulas models with different choices of linking copulas as the number of categories K increases. For a small K dependence in the tails cannot be easily quantified. Hence, for example, when the true copula is the t_5 which has the same upper and lower tail dependence, the algorithm selected either t_5 or BVN which has zero lower and upper tail dependence, because both copulas provide reflection symmetric dependence.

To check whether the χ^2_{s-q} is a good approximation for the distribution of the M_2 statistic under the null hypothesis, samples of sizes 500 and 1000 were generated

Table 3.7: Small sample of size $n = 500$ simulations (10^3 replications) from the bi-factor and second-order factor models with various linking copulas and frequencies of the true bivariate copula identified using the model selection algorithm.

Bi-factor	Model 1			Model 2			Model 3			Model 4		
	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$
X_0	Gumbel	992	1000	t_5	984	1000	Gumbel	996	1000	t_5	975	1000
X_1	Gumbel	858	956	t_5	597	806	t_5	585	789	Gumbel	888	958
X_2	Gumbel	870	951	t_5	588	799	t_5	569	775	Gumbel	894	969
X_3	Gumbel	846	950	t_5	546	777	s.Gumbel	844	945	s.Gumbel	865	947
X_4	Gumbel	844	942	t_5	589	805	s.Gumbel	878	949	s.Gumbel	900	956
Second-order	Model 1			Model 2			Model 3			Model 4		
	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$	Copula	$K=3$	$K=5$
X_0	Gumbel	901	848	t_5	664	819	Gumbel	892	987	t_5	648	765
X_1	Gumbel	895	975	t_5	735	939	t_5	756	933	Gumbel	918	990
X_2	Gumbel	892	962	t_5	686	911	t_5	705	910	Gumbel	918	991
X_3	Gumbel	891	981	t_5	711	915	s.Gumbel	901	980	s.Gumbel	902	982
X_4	Gumbel	900	984	t_5	743	926	s.Gumbel	904	984	s.Gumbel	919	980

from various bi-factor second-order copula models. Table 3.8 contains four common nominal levels of the M_2 statistic under the bi-factor and second-order copula models with different bivariate copulas. As can be seen in the table the observed levels of M_2 are close to the nominal α levels and remain accurate even for extremely sparse tables ($d = 16$ and $K = 5$).

Table 3.8: Small sample of size $n = \{500, 1000\}$ simulations (10^3 replications) from bi-factor and second-order copula models and the empirical rejection levels at $\alpha = \{0.20, 0.10, 0.05, 0.01\}$, degrees of freedom (df), mean and variance.

			M_2						
Copula	n	K	df	Mean	Variance	α=0.20	α=0.10	α=0.05	α=0.01
Bi-factor copula model									
BVN	500	3	448	449.0	912.8	0.206	0.100	0.060	0.016
		5	1888	1885.5	4858.3	0.210	0.117	0.065	0.024
	1000	3	448	448.7	879.0	0.192	0.097	0.051	0.020
		5	1888	1886.5	4332.5	0.202	0.108	0.064	0.015
Gumbel	500	3	448	449.9	887.3	0.216	0.111	0.053	0.011
		5	1888	1886.6	4709.7	0.225	0.126	0.070	0.015
	1000	3	448	448.9	864.0	0.201	0.102	0.050	0.015
		5	1888	1888.6	4332.1	0.226	0.107	0.069	0.014
t_5	500	3	448	448.7	907.3	0.202	0.088	0.048	0.018
		5	1888	1886.6	4479.4	0.204	0.107	0.053	0.017
	1000	3	448	448.6	834.9	0.184	0.090	0.050	0.014
		5	1888	1890.3	4008.5	0.218	0.103	0.052	0.015
Second-order copula model									
BVN	500	3	460	462.2	1001.2	0.220	0.113	0.055	0.016
		5	1900	1903.5	3736.2	0.214	0.112	0.052	0.010
	1000	3	460	461.3	1023.9	0.220	0.109	0.064	0.013
		5	1900	1906.5	3918.2	0.230	0.130	0.068	0.012
Gumbel	500	3	460	464.5	1011.3	0.233	0.117	0.073	0.024
		5	1900	1909.2	5099.8	0.245	0.129	0.064	0.008
	1000	3	460	461.9	871.2	0.203	0.106	0.049	0.009
		5	1900	1908.5	3977.0	0.239	0.129	0.067	0.015
t_5	500	3	460	465.3	1362.4	0.247	0.145	0.091	0.039
		5	1900	1904.7	3740.6	0.226	0.113	0.050	0.010
	1000	3	460	461.8	900.1	0.214	0.108	0.055	0.010
		5	1900	1908.1	3864.9	0.229	0.131	0.072	0.015

3.6 Application

Alexithymia is a personality construct that is defined as a difficulty identifying, experiencing or describing emotions (Schroeders et al., 2021). The most utilized measure of alexithymia in empirical research is the Toronto Alexithymia Scale (Bagby et al. 1994; Gignac et al. 2007; Tuliao et al. 2020). It is composed of $d = 20$ items that can be subdivided into $G = 3$ non-overlapping groups: $d_1 = 7$ items to assess difficulty identifying feelings (DIF), $d_2 = 5$ items to assess difficulty describing feelings (DDF) and $d_3 = 8$ items to assess externally oriented thinking (EOT). We use a dataset of 1925 university students from the French-speaking region of Belgium (Briganti and Linkowski, 2020). Students were 17 to 25 years old and 58% of them were female and 42% were male. They were asked to respond to each item using one of $K = 5$ categories: "$1 =$ completely disagree", "$2 =$ disagree", "$3 =$ neutral, "$4 =$ agree", "$5 =$ completely agree". The dataset and full description of the items can be found in Table 3.9 and the R package **BGGM** (Williams and Mulder, 2020).

For these items, a respondent might be thinking about the average "sensation" of many past relevant events, leading to latent means. That is, based on the item descriptions, this seems more natural than a discretized maxima or minima. Since the sample is a mixture (male and female students) we can expect a priori that a bi-factor or second-order copula model with t_ν copulas might be plausible, as in this case the items can be considered as mixtures of discretized means.

In Table 3.10 we summarize the averages of polychoric semi-correlations for all pairs within each group and for all pairs of items along with the theoretical semi-correlations in Section 2.1.1 under different choices of copulas. For a BVN/t_ν copula the copula parameter is the sample polychoric correlation, while for a Gum-

Table 3.9: The Toronto Alexithymia Scale with 20 items categorized into 3 groups.

Number	Item	Group
1	I am often confused about what emotion I am feeling	Difficulty identifying feelings
2	It is difficult for me to find the right words for my feelings	Difficulty describing feelings
3	I have physical sensations that even doctors don't understand	Difficulty identifying feelings
4	I am able to describe my feelings easily	Difficulty describing feelings
5	I prefer to analyze problems rather than just describe them	Externally oriented thinking
6	When I am upset, I don't know if I am sad, frightened, or angry	Difficulty identifying feelings
7	I am often puzzled by sensations in my body	Difficulty identifying feelings
8	I prefer to just let things happen rather than to understand why they turned out that way	Externally oriented thinking
9	I have feelings that I can't quite identify	Difficulty identifying feelings
10	Being in touch with emotions is essential	Externally oriented thinking
11	I find it hard to describe my feelings more	Difficulty describing feelings
12	People tell me to describe my feelings more	Difficulty describing feelings
13	I don't know what's going on inside me	Difficulty identifying feelings
14	I often don't know why I am angry	Difficulty identifying feelings
15	I prefer talking to people about their daily activities rather than their feelings	Externally oriented thinking
16	I prefer to watch "light" entertainment shows rather psychological dramas	Externally oriented thinking
17	It is difficult for me to reveal my innermost feelings, even to close friends	Difficulty describing feelings
18	I can feel close to someone, even in moments of silence	Externally oriented thinking
19	I find examination of my feelings useful in solving personal problems	Externally oriented thinking
20	Looking for hidden meanings in movies or plays distracts from their enjoyment	Externally oriented thinking

Table 3.10: Average observed polychoric correlations and semi-correlations for all pairs within each group and for all pairs of items for the Toronto Alexithymia Scale (TAS), along with the corresponding theoretical semi-correlations for BVN, t_5, Frank, Gumbel , and survival Gumbel (s.Gumbel) copulas.

	All items			Items in group 1			Items in group 2			Items in group 3		
	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+	ρ_N	ρ_N^-	ρ_N^+
Observed	0.17	0.21	0.20	0.34	0.36	0.29	0.42	0.37	0.40	0.19	0.26	0.29
BVN	0.17	0.07	0.07	0.34	0.16	0.16	0.42	0.21	0.21	0.19	0.08	0.08
t_5	0.17	0.23	0.23	0.34	0.31	0.31	0.42	0.35	0.35	0.19	0.24	0.24
Frank	0.17	0.04	0.04	0.34	0.10	0.10	0.42	0.13	0.13	0.19	0.05	0.05
Gumbel	0.17	0.05	0.22	0.34	0.11	0.37	0.42	0.14	0.43	0.19	0.05	0.24
s.Gumbel	0.17	0.22	0.05	0.34	0.37	0.11	0.42	0.43	0.14	0.19	0.24	0.05

bel/s.Gumbel copula the polychoric correlation was converted to Kendall's tau with the relation in (3.6) and then from Kendall's τ to Gumbel/s.Gumbel copula parameter via the functional inverse in (4.10). The summary of findings from the diagnostics in the table show that:

- for the first group of items there is more probability in the joint lower tail suggesting s.Gumbel linking copulas to join each item in this group with the DIF factor;
- for the second group of items there is more probability in the joint lower and upper tail suggesting t_ν linking copulas to join each item in this group with the DDF factor;
- for the third group of items there is more probability in the joint lower and upper tail suggesting t_ν linking copulas to join each item in this group with the EOT factor;
- for the items overall there is more probability in the joint lower and upper tail suggesting t_ν linking copulas to join each item or group specific factor (second-order model) with the common factor.

Hence, a bi-factor or second-order copula model with the aforementioned linking copulas might provide a better fit that the (Gaussian) models with BVN copulas.

Then, we fit the bi-factor and second-order models with the copulas selected by the heuristic algorithm in Section 3.3.1. For a baseline comparison, we also fit their special cases; these are the one- and two-factor copula models where we have also selected the bivariate copulas using the heuristic algorithm proposed by Kadhem and Nikoloulopoulos (2021b) which is presented in Section 2.3 from previous chapter. To

show the improvement of the copula models over their Gaussian analogues, we have also fitted all the classes of copula models with BVN copulas. The fitted models are compared via the AIC, since the number of parameters is not the same between the models. In addition, we use Vuong's test (Vuong, 1989) to show if (a) the best fitted model according to the AICs provides better fit than the other fitted models and (b) a model with the selected copulas provides better fit than the one with BVN copulas. The Vuong test is the sample version of the difference in Kullback-Leibler divergence between two models and can be used to differentiate two parametric models which could be non-nested. For the Vuong's test we provide the 95% confidence interval of the test statistic (Joe, 2014, page 258). If the interval does not contain 0, then the best fitted model according to the AICs is better if the interval is completely above 0. To assess the overall goodness-of-fit of the bi-factor and second-order copula models, we use the M_2 statistic (Maydeu-Olivares and Joe, 2006).

Table 3.11 gives the AICs, the 95% CIs of Vuong's tests and the M_2 statistics for all the fitted models. The best fitted bi-factor copula model results when we use s.Gumbel for the DIF factor, t_3 for both the DDF and EOT factors and t_2 for the common factor (alexithymia). This is in line with the preliminary analyses based on the interpretations of items as mixtures of means and the diagnostics in Table 3.10. It is revealed that the DIF items and DIF factor are discretized and latent minima, respectively, as the participants seem to reflect that they "disagree" or "completely disagree" having difficulty identifying feelings. From the Vuong's 95% Cls and M_2 statistics it is shown that factor copula models provide a big improvement over their Gaussian analogues and that the selected bi-factor copula model outperforms all the fitted models.

Table 3.11: AICs, Vuong's 95% CIs, and M_2 statistics for the 1-factor, 2-factor, bi-factor and second-order copula models with BVN copulas and selected copulas, along with the maximum deviations of observed and expected counts for all pairs within each group and for all pairs of items for the Toronto Alexithymia Scale.

	1-factor		2-factor		Bi-factor		Second-order	
	BVN	Selected	BVN	Selected	BVN	Selected	BVN	Selected
AIC	107135.8	105504.0	106189.5	103893.5	105507.7	103200.9	105878.6	104133.7
Vuong's 95% CI[a]	(0.35,0.50)		(0.53,0.69)		(0.51,0.69)		(0.38,0.52)	
Vuong's 95% CI[b]	(0.93,1.13)	(0.55,0.67)	(0.69,0.88)	(0.13,0.23)	(0.51,0.69)		(0.61,0.80)	(0.21,0.29)
M_2	14723.8	9865.0	9195.7	7383.7	11664.7	6381.5	13547.1	7341.2
df	3020	3020	3001	3000	3000	3000	3017	3017
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Maximum discrepancy								
Items in Group 1	71	63	71	60	69	55	70	61
Items in Group 2	112	98	113	83	77	48	84	55
Items in Group 3	87	74	81	52	80	45	82	53
All items	112	98	113	83	80	55	84	61

[a] Selected factor copula model versus its Gaussian special case.
[b] Selected Bi-factor copula model versus any other fitted model.

Although the selected bi-factor copula model shows substantial improvement over the Gaussian bi-factor model or any other fitted model, it is not so clear from the goodness-of-fit p-values that the response patterns are satisfactorily explained by using the linking copulas selected by the heuristic algorithm. This is not surprising since one should expect discrepancies between the postulated parametric model and the population probabilities, when the sample size or dimension is sufficiently large (Maydeu-Olivares and Joe, 2014). To further show that the fit has been improved we have calculated the maximum deviations of observed and model-based counts for each bivariate margin, that is, $D_{j_1 j_2} = n \max_{y_1,y_2} |p_{j_1,j_2,y_1,y_2} - \pi_{j_1,j_2,y_1,y_2}(\hat{\boldsymbol{\theta}})|$. In Table 3.11 we summarize the averages of these deviations for all pairs within each group and for all pairs of items. Overall, the maximum discrepancies have been sufficiently reduced in the selected bi-factor model.

Table 3.12 gives the copula parameter estimates in Kendall's τ scale and their standard errors (SEs) for the selected bi-factor copula model and the Gaussian bi-factor model as the benchmark model. The SEs of the estimated parameters are obtained by the inversion of the Hessian matrix at the second step of the IFM method. These SEs are adequate to assess the flatness of the log-likelihood. Proper SEs that account for the estimation of cutpoints can be obtained by jackknifing the two-stage estimation procedure. The loading parameters ($\hat{\tau}$'s converted to BVN copula parameters via the functional inverse in (3.6) and then to loadings using the relations in Section 3.1.3) show that the common alexithymia factor is mostly loaded on DIF and DDF items, suggesting that items in the domains DIF and DDF are good indicators for alexithymia. The items in the EOT although they loaded on the EOT latent factor, they had poor loadings in the common alexithymia factor.

Table 3.12: Estimated copula parameters and their standard errors (SE) in Kendall's τ scale for the Bi-factor copula models with BVN copulas and selected copulas for the Toronto Alexithymia Scale.

	Bi-factor copula model with BVN copulas				Bi-factor copula model with selected copulas					
	Common factor		Group-specific factors		Common factor			Group-specific factors		
Items	Est	SE	Est	SE	Copulas	Est	SE	Copulas	Est	SE
1	0.42	0.01	0.23	0.02	t_2	0.49	0.02	s.Gumbel	0.09	0.03
3	0.14	0.02	0.24	0.02	t_2	0.16	0.02	s.Gumbel	0.37	0.02
6	0.22	0.02	0.29	0.02	t_2	0.29	0.02	s.Gumbel	0.23	0.02
7	0.11	0.02	0.31	0.02	t_2	0.09	0.02	s.Gumbel	0.53	0.04
9	0.38	0.01	0.34	0.02	t_2	0.47	0.02	s.Gumbel	0.24	0.02
13	0.36	0.01	0.46	0.02	t_2	0.49	0.02	s.Gumbel	0.32	0.03
14	0.21	0.02	0.36	0.02	t_2	0.30	0.02	s.Gumbel	0.27	0.03
2	0.71	0.02	-0.24	0.10	t_2	0.46	0.02	t_3	0.53	0.02
4	0.55	0.01	0.02	0.04	t_2	0.41	0.02	t_3	0.58	0.03
11	0.35	0.01	0.13	0.03	t_2	0.33	0.02	t_3	0.20	0.03
12	0.34	0.02	0.29	0.04	t_2	0.29	0.02	t_3	0.23	0.03
17	0.31	0.02	0.38	0.06	t_2	0.24	0.02	t_3	0.25	0.03
5	0.06	0.02	0.33	0.02	t_2	0.10	0.02	t_3	0.34	0.02
8	0.11	0.02	0.30	0.02	t_2	0.16	0.02	t_3	0.33	0.02
10	0.12	0.02	0.27	0.02	t_2	0.14	0.02	t_3	0.30	0.02
15	0.15	0.02	0.19	0.02	t_2	0.12	0.02	t_3	0.19	0.02
16	0.03	0.02	0.23	0.02	t_2	0.03	0.02	t_3	0.24	0.02
18	-0.02	0.02	0.28	0.02	t_2	0.03	0.02	t_3	0.29	0.02
19	0.07	0.02	0.40	0.02	t_2	0.10	0.02	t_3	0.43	0.02
20	0.06	0.02	0.27	0.02	t_2	0.10	0.02	t_3	0.26	0.02

3.7 Software

R functions for estimation, simulation, model selection and goodness-of-fit of the bi-factor and second-order copula models are part of the R package **FactorCopula** (Kadhem and Nikoloulopoulos, 2021c). All the analyses presented in Section 3.6 are given as code examples in the package.

3.8 Chapter summary

We have proposed bi-factor and second-order copula models for item response data that can be split into non-overlapping groups. Our copula constructions include the Gaussian bi-factor and second-order models as special cases when we construct the proposed models with BVN copulas. They also provide substantial improvement over the Gaussian and other competing models (one and two factor copula model with BVN and selected copulas) based on AIC, Vuong's and M_2 goodness-of-fit statistics. This improvement relies on the fact that when we use appropriate bivariate copulas other than BVN copulas in the construction, there is an interpretation of latent variables that can be maxima/minima or mixture of means instead of means.

Chapter 4

Factor tree copula models for item response data

The factor copula models in Chapter 2 require the conditional independence assumption, where the observed variables are conditionally independent given some latent variables. This assumption implies that the dependence among the observed variables is adequately explained by those latent variables. However, this assumption might not be realistic in some scenarios. For example, violation of the conditional independence assumption can occur if we have items that can be split into non-overlapping groups. To alleviate the violation of this assumption, a possibility is to use the bi-factor and second-order copulas proposed in Chapter 3 to model dependencies between and within different groups.

In this chapter, without a priori knowledge of the subgroups of items, we extend the factor copula models in Chapter 2 for item response data to model the residual dependence. The main new contribution in this chapter is the construction of factor copula models with conditional dependence structure, where we combine the factor

copula models with an 1-truncated vine copula for item response data. These models introduce conditional dependence structure given very few latent variables.

The proposed models are built based on arbitrary bivariate parametric copula families and thus allow for a flexible vine structure. Bivariate copulas other than BVN can be called to model tail asymmetry/dependence in the data. In order to build plausible models, accounting for different tail behaviour, we propose model selection algorithms that select a suitable vine structure and bivariate parametric copulas. Hereafter, we will refer to the model as factor tree copula.

We illustrate the proposed methodology by re-analysing a real dataset. We show that the factor tree copula models with the selected vine tree and copulas (obtained from the model selection algorithms) provide a substantial improvement over relevant benchmark models.

The rest of the chapter is as follows. In Section 4.1, we introduce the combined factor/truncated vine copula models for item response data. Section 4.2 provides estimation techniques and computational details. Section 4.3 discusses vine tree and bivariate copula selection. Section 4.4 has an extensive simulation study to assess the estimation techniques and model selection algorithms. Our methodology is illustrated using real data in Section 4.5, followed with a summary in Section 4.6.

4.1 Factor tree copula models for item response

This section introduces the theory of the combined factor/truncated vine copula models for item response data. Before that, the first two sections provide some background about factor (Nikoloulopoulos and Joe, 2015) and truncated vine (Panagiotelis et al., 2012, 2017) copula models for discrete responses.

4.1.1 Factor copula models

We first introduce the notation used in this chapter. Let $\mathbf{Y} = \{Y_1, \ldots, Y_d\}$ denote the vector with the item response variables that are all measured on an ordinal scale; $Y_j \in \{0, \ldots, K_j - 1\}$. Let the cutpoints in the uniform $U(0,1)$ scale for the jth item be $a_{j,k}$, $k = 1, \ldots, K-1$, with $a_{j,0} = 0$ and $a_{j,K} = 1$. These correspond to $a_{j,k} = \Phi(\alpha_{j,k})$, where $\alpha_{j,k}$ are cutpoints in the normal $N(0,1)$ scale.

The p-factor model assumes that $\mathbf{Y}$, with corresponding realizations $\mathbf{y} = \{y_1, \ldots, y_d\}$, is conditionally independent given the p-dimensional latent vector $\mathbf{X} = (X_1, \ldots, X_p)$. The joint probability mass function (pmf) of the p-factor model is

$$\begin{aligned} \pi_d(\mathbf{y}) &= \Pr(Y_1 = y_1, \ldots, Y_d = y_d) \\ &= \int \prod_{j=1}^{d} \Pr(Y_j = y_j | X_1 = x_1, \ldots, X_p = x_p)\, dF_{\mathbf{X}}(x), \end{aligned} \tag{4.1}$$

where $F_{\mathbf{X}}$ is the distribution of the latent vector $\mathbf{X}$. The factor copula methodology uses a set of bivariate copulas that link the items to the latent variables to specify $\Pr(Y_j = y_j | X_1 = x_1, \ldots, X_p = x_p)$. Below we include the theory for one and two factors.

For the 1-factor model, let X_1 be a latent variable that is standard uniform. From Sklar (1959), there is a bivariate copula $C_{X_1 j}$ such that $\Pr(X_1 \le x, Y_j \le y) = C_{X_1 j}\big(x, F_j(y)\big)$ for $0 \le x \le 1$ where $F_j(y) = a_{j,y+1}$ is the cdf of Y_j. Then it follows that

$$F_{j|X_1}(y|x) := \Pr(Y_j \le y | X_1 = x) = \frac{\partial C_{X_1 j}(x, a_{j,y+1})}{\partial x} = C_{j|X_1}(a_{j,y+1}|x). \tag{4.2}$$

Hence, the pmf for the 1-factor copula model becomes

$$\pi_d(\mathbf{y}) \;=\; \int_0^1 \prod_{j=1}^{d} \Pr(Y_j = y_j | X_1 = x)\, dx = \int_0^1 \prod_{j=1}^{d} f_{j|X_1}(y_j|x)\, dx,$$

where

$$f_{j|X_1}(y|x) = C_{j|X_1}(a_{j,y+1}|x) - C_{j|X_1}(a_{j,y}|x). \tag{4.3}$$

For the 2-factor copula model, let X_1, X_2 be latent variables that are independent uniform $U(0,1)$ random variables. Let $C_{X_1 j}$ be defined as in the 1-factor copula model and $C_{X_2 j}$ be a bivariate copula such that

$$\Pr(X_2 \le x_2, Y_j \le y | X_1 = x_1) = C_{X_2 j}\big(x_2, F_{j|X_1}(y|x_1)\big),$$

where $F_{j|X_1}$ is given in (4.2). Then for $0 \le x_1, x_2 \le 1$,

$$\begin{aligned} F_{X_2 j|X_1}(x_2,&y|x_1) := \Pr(Y_j \le y | X_1 = x_1, X_2 = x_2) \\ &= \frac{\partial}{\partial x_2} \Pr(X_2 \le x_2, Y_j \le y | X_1 = x_1) = \frac{\partial}{\partial x_2} C_{X_2 j}\Big(x_2, F_{j|X_1}(y|x_1)\Big) \\ &= C_{j|X_2}\Big(F_{j|X_1}(y|x_1)|x_2\Big). \end{aligned} \tag{4.4}$$

Hence, the pmf for the 2-factor copula model is

$$\begin{aligned} \pi_d(\mathbf{y}) \;&=\; \int_0^1 \int_0^1 \prod_{j=1}^{d} \Pr(Y_j = y_j | X_1 = x_1, X_2 = x_2)\, dx_1 dx_2 \\ &=\; \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2 j|X_1}\big(x_2, y_j|x_1\big)\, dx_1 dx_2, \end{aligned}$$

where

$$f_{X_2 j|X_1}(x_2, y|x_1) = C_{j|X_2}\Big(F_{j|X_1}(y|x_1)|x_2\Big) - C_{j|X_2}\Big(F_{j|X_1}(y-1|x_1)|x_2\Big). \tag{4.5}$$

4.1.2 1-truncated vine copula models

Vine copula models involve $d-1$ trees, the first tree represents dependence (as edges) amongst d variables (as nodes). Then the edges become nodes in the next tree, involving the conditional dependencies given a common variable. This process continues until tree $d-1$ that includes two nodes and one edge, representing conditional dependence of two variables given $d-2$ variables (Chang and Joe, 2019).

If one is restricted to the first tree, that is truncation at level 1, then the result is a Markov tree dependence structure where two variables not connected by an edge are conditionally independent given the variables in the tree between them. In a Markov tree or 1-truncated vine with d variables, $d-1$ of the $d(d-1)/2$ possible pairs are identified as the edges of a tree with d nodes corresponding to the items, i.e., there are a total of $d-1$ edges, where two connected pairs of items form an edge. Let j and k be indices for any pairs of items with $1 \leq k < j \leq d$. For a given vine tree structure, let $\mathcal{E}$ denote the set of edges. Each edge of $jk \in \mathcal{E}$ is represented with a bivariate copula C_{jk} such that

$$\Pr(Y_j \leq y_j, Y_k \leq y_k) = C_{jk}\big(F_j(y_j), F_k(y_k)\big) = C_{jk}(a_{j,y_j+1}, a_{k,y_k+1}).$$

Since the densities of vine copulas can be factorized in terms of bivariate linking copulas and lower-dimensional margins, they are computationally tractable for high-dimensional continuous variables. Nevertheless, the cdf of d-dimensional vine copula lacks a closed form and requires $(d-1)$-dimensional integration (Joe, 1997). Hence, in order to derive the d-dimensional pmf using finite differences of the d-dimensional cdf (e.g., Braeken et al. 2007 or Nikoloulopoulos 2013a) poses non-negligible numerical challenges. This problem has been solved by Panagiotelis et al. (2012) who

decomposed the d-dimensional pmf into finite differences of bivariate copula cdfs. Hence, the pmf of an 1-truncated vine model takes the form

$$\pi_d(\mathbf{y}) = \prod_{j=1}^{d} \Pr(Y_j = y_j) \prod_{jk \in \mathcal{E}} \frac{\Pr(Y_j = y_j, Y_k = y_k)}{\Pr(Y_j = y_j) \Pr(Y_k = y_k)}, \tag{4.6}$$

where $\Pr(Y_j = y_j, Y_k = y_k) = C_{jk}(a_{j,y_j+1}, a_{k,y_k+1}) - C_{jk}(a_{j,y_j}, a_{k,y_k+1}) - C_{jk}(a_{j,y_j+1}, a_{k,y_k}) + C_{jk}(a_{j,y_j}, a_{k,y_k})$ and $\Pr(Y = y) = a_{j,y+1} - a_{j,y}$.

4.1.3 Combined factor/truncated vine copula models

In this section we combine the factor copula model with an 1-truncated vine copula to account for the residual dependence. The pmf of an 1-truncated vine copula in (4.6) can be used in the pmf of the factor copula model in (4.1) instead of the product to capture any residual dependencies. Hence the pmf of the combined factor/truncated vine copula model takes the form

$$\pi_d(\mathbf{y}) = \int \prod_{j=1}^{d} \Pr\left(Y_j = y_j | \mathbf{X} = \mathbf{x}\right) \times \prod_{[jk] \in \mathcal{E}} \frac{\Pr\left(Y_j = y_j, Y_k = y_k | \mathbf{X} = \mathbf{x}\right)}{\Pr\left(Y_j = y_j | \mathbf{X} = \mathbf{x}\right) \Pr\left(Y_k = y_k | \mathbf{X} = \mathbf{x}\right)} \, dF_{\mathbf{X}}(\mathbf{x}).$$

With one factor and an 1-truncated vine given the latent variable X_1 (hereafter 1-factor tree) let $C_{jk;X_1}$ be a bivariate copula such that

$$\Pr(Y_j \leq y_j, Y_k \leq y_k | X_1 = x_1) = C_{jk;X_1}\big(F_{j|X_1}(y_j|x_1), F_{k|X_1}(y_k|x_1)\big),$$

where $F_{j|X_1}$ and $F_{k|X_1}$ are given in (4.2). Then for a given 1-truncated vine structure with a set of edges $\mathcal{E}$, the pmf of the 1-factor tree copula model is

$$\pi_d(\mathbf{y}) = \int_0^1 \prod_{j=1}^{d} f_{j|X_1}(y_j|x) \prod_{jk\in\mathcal{E}} \frac{f_{jk|X_1}(y_j, y_k|x_1)}{f_{j|X}(y_j|x)\, f_{k|X}(y_k|x)}\, dx, \tag{4.7}$$

where

$$f_{jk|X_1}(y_j, y_k|x_1) = C_{jk|X_1}\big(F^+_{j|X_1}, F^+_{k|X_1}\big) - C_{jk|X_1}\big(F^-_{j|X_1}, F^+_{k|X_1}\big)$$
$$- C_{jk|X_1}\big(F^+_{j|X_1}, F^-_{k|X_1}\big) + C_{jk|X_1}\big(F^-_{j|X_1}, F^-_{k|X_1}\big)$$

and $f_{j|X}(y_j|x)$, $f_{k|X}(y_k|x)$ are given in (4.3). In the above $F^+_{j|X_1} = F_{j|X_1}(y|x)$ and $F^-_{j|X_1} = F_{j|X_1}(y-1|x)$.

Figure 4.1 depicts the graphical representation of a 1-factor tree copula model with $d = 5$ items as a 2-truncated vine. Tree 1 shows the typical 1-factor model, while Tree 2 accounts for the residual dependence by the pairwise conditional dependencies of two items conditioned on the factor X_1.

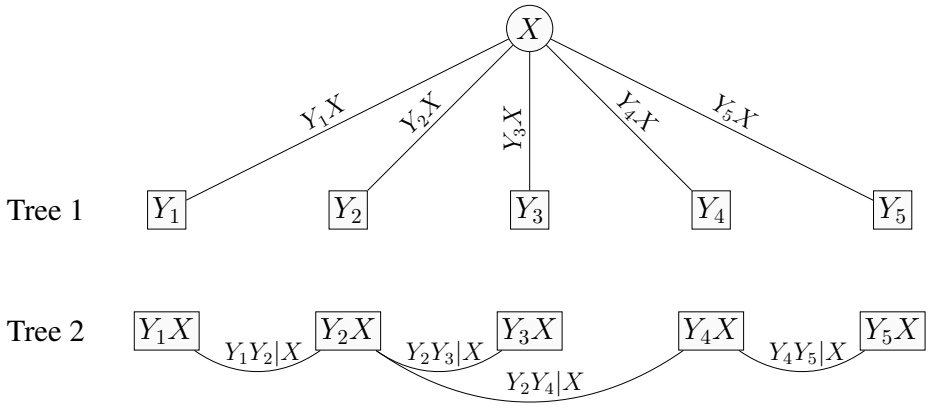


Figure 4.1: Graphical representation of a 1-factor tree copula model with $d = 5$ items. The first tree is the 1-factor model. The residual dependence is captured in Tree 2 with an 1-truncated vine model.

With two factors and an 1-truncated vine given the latent variables X_1, X_2 (hereafter 2-factor tree), let $C_{jk;X_1,X_2}$ be a bivariate copula cdf such that

$$\Pr(Y_j \leq y_j, Y_k \leq y_k | X_1, X_2) = C_{jk;X_1X_2}\big(F_{X_2j|X_1}(x_2, y_j|x_1), F_{X_2k|X_1}(x_2, y_k|x_1)\big),$$

where $F_{X_2j|X_1}$ and $F_{X_2k|X_1}$ are given in (4.4). Then for a given vine structure with a set of edges $\mathcal{E}$, the pmf of the 2-factor tree copula model is

$$\pi_d(\mathbf{y}) = \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2j|X_1}\left(x_2, y_j|x_1\right) \times \prod_{[jk]\in\mathcal{E}} \frac{f_{jk|X_1X_2}(y_j, y_k|x_1, x_2)}{f_{X_2j|X_1}\left(x_2, y_j|x_1\right) f_{X_2k|X_1}\left(x_2, y_k|x_1\right)} \, dx_1 dx_2, \qquad (4.8)$$

where

$$\begin{aligned} f_{jk|X_1X_2}(y_j,y_k|x_1, x_2) = C_{jk|X_1X_2}\big(F^+_{X_2j|X_1}, F^+_{X_2k|X_1}\big) - C_{jk|X_1X_2}\big(F^-_{X_2j|X_1}, F^+_{X_2k|X_1}\big) \\ - C_{jk|X_1X_2}\big(F^+_{X_2j|X_1}, F^-_{X_2k|X_1}\big) + C_{jk|X_1X_2}\big(F^-_{X_2j|X_1}, F^-_{X_2k|X_1}\big). \end{aligned}$$

and $f_{X_2j|X_1}(x_2, y_j|x_1)$, $f_{X_2k|X_1}(x_2, y_k|x_1)$ are as in (4.5). In the above $F^+_{X_2j|X_1} = F_{X_2j|X_1}(x_2, y|x_1)$ and $F^-_{X_2j|X_1} = F_{X_2j|X_1}(x_2, y-1|x_1)$.

Figure 4.2 depicts the graphical representation of a 2-factor tree copula model with $d = 5$ items as a 3-truncated vine. Trees 1 and 2 show the common 2-factor model, while Tree 3 involves the pairwise conditional dependencies of two items given the factors.

For parametric 1-factor and 2-factor tree copula models, we let C_{X_1j}, C_{X_2j} and $C_{jk;\mathbf{X}}$ be parametric bivariate copulas, say with parameters θ_{1j}, θ_{2j}, and δ_{jk}, respectively. For the set of all parameters, let $\boldsymbol{\theta} = \{a_{jk}, \theta_{1j}, \delta_{jk} : j = 1, \ldots, d; k =$

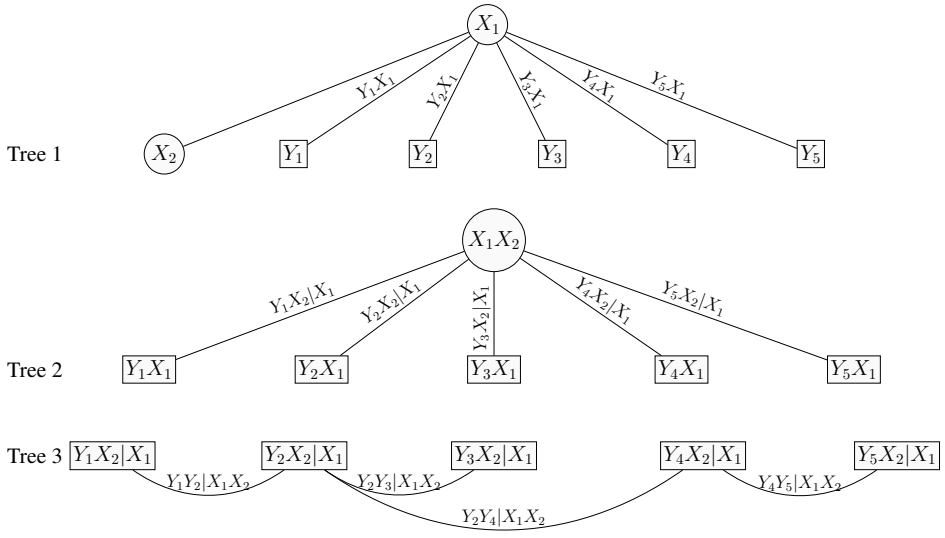


Figure 4.2: Graphical representation of a 2-factor tree copula model with $d = 5$ items. The first and second trees represent the 2-factor model. The residual dependence is captured in Tree 3 with an 1-truncated vine model. Note that the factors are linked to one another with an independent copula in Tree 1.

$1, \ldots, K-1; jk \in \mathcal{E}\}$ for the 1-factor tree copula model and $\boldsymbol{\theta} = \{a_{jk}, \theta_{1j}, \theta_{2j}, \delta_{jk} : j = 1, \ldots, d; k = 1, \ldots, K-1; jk \in \mathcal{E}\}$ for the 2-factor tree copula model.

4.1.4 Choices of parametric bivariate copulas

In line with Nikoloulopoulos and Joe (2015), we use bivariate parametric copulas that can be used when considering latent maxima, minima or mixtures of means. For different dependent items based on latent maxima or minima, multivariate extreme value and copula theory (e.g., Joe 1997) can be used to select suitable copulas that link observed to latent variables. Copulas that arise from extreme value theory have more probability in one joint tail (upper or lower) than expected with a discretized MVN distribution or a MVN copula with discrete margins. If item responses are based on discretizations of latent variables that are means, then it is possible that there can be more probability in both the joint upper and joint lower tail, compared

with discretized MVN models. This happens if the respondents consist of a 'mixture' population (e.g., different locations or genders). From the theory of elliptical distributions and copulas (e.g., McNeil et al. 2005), it is known that the multivariate Student-t distribution as a scale mixture of MVN has more dependence in the tails. Extreme value and elliptical copulas can model item response data that have reflection asymmetric and symmetric dependence, respectively.

Choices of copulas with upper or lower tail dependence are better if the items have more probability in joint lower or upper tail than would be expected with the BVN copula. We provide below the bivariate copula choices we consider:

- A model with BVN copulas has latent (ordinal) variables that can be considered as (discretized) means and and there is less probability in both the joint upper and joint lower tail as the BVN copula has reflection symmetry and tail independence.

- A model with bivariate Gumbel copulas has latent (ordinal) variables that can be considered as (discretized) maxima and there is more probability in the joint upper tail as the Gumbel copula has reflection asymmetry and upper tail dependence.

- A model with bivariate survival Gumbel copulas has latent (ordinal) variables that can be considered as (discretized) minima and there is more probability in the joint lower tail as the survival Gumbel copula has reflection asymmetry and lower tail dependence.

- A model with bivariate t_ν copulas has latent (ordinal) variables that can be considered as mixtures of (discretized) means, since the bivariate Student-t distribution arises as a scale mixture of bivariate normals. A small value of

ν, such as $1 \leq \nu \leq 5$, leads to a model with more probabilities in the joint upper and joint lower tails compared with the BVN copula as the t_ν copula has reflection symmetric upper and lower tail dependence.

For the residual part of the model in addition to the aforementioned bivariate parametric copulas for computational improvements we can use the Archimedean Frank copula. For all the bivariate margins to have more probability in the joint lower or upper tail, it only suffices that the bivariate copulas in the first trees (factor part) to have upper/lower tail dependence and is not necessary for the bivariate copulas in the higher trees (residual part) to have tail dependence. For discrete data, such as item response, the Frank copula has the same tail behaviour with the BVN copula but provides simplified computations as it has a closed from cdf and thus it can preferred over the BVN copula for the residual part of the model that involves finite differences of bivariate copula cdfs.

4.2 Estimation

With sample size n and data $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the joint log-likelihood of the factor tree copula models is

$$\ell(\boldsymbol{\theta}; \mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^{n} \log \pi_d(\mathbf{y}_i; \boldsymbol{\theta}), \tag{4.9}$$

with $\pi_d(\mathbf{y})$ as defined in (4.7) and (4.8) for the 1-factor and 2-factor tree copula model, respectively. Maximization of (4.9) is numerically possible but time-consuming for large d because of many univariate cutpoints and dependence parameters. Hence, we approach estimation using the two-step IFM method proposed by Joe (2005).

In the first step, the cutpoints are estimated using the univariate sample proportions. The univariate cutpoints for the jth item are estimated as $\hat{a}_{j,k} = \sum_{y=0}^{k} p_{j,y}$, where $p_{j,y}$, $y = 0, \ldots, K-1$ for $j = 1, \ldots, d$ are the univariate sample proportions. In the second step of the IFM method, the joint log-likelihood in (4.9) is maximized over the copula parameters with the cutpoints fixed as estimated at the first step. The estimated copula parameters can be obtained by using a quasi-Newton (Nash, 1990) method applied to the logarithm of the joint likelihood.

For the 1-factor tree copula model, numerical evaluation of the joint pmf can be achieved with the following steps:

1. Calculate Gauss-Legendre quadrature (Stroud and Secrest, 1966) points $\{x_q : q = 1, \ldots, n_q\}$ and weights $\{w_q : q = 1, \ldots, n_q\}$ in terms of standard uniform.

2. Numerically evaluate the joint pmf in (4.7) via the following approximation:

$$\sum_{q=1}^{n_q} w_q \prod_{j=1}^{d} f_j(y_j|x_q) \prod_{[jk]\in\mathcal{E}} \frac{f_{jk|X_1}(y_j, y_k|x_q)}{f_{j|X}(y_j|x_q) f_{k|X}(y_k|x_q)}.$$

For the 2-factor tree copula model, numerical evaluation of the joint pmf can be achieved with the following steps:

1. Calculate Gauss-Legendre quadrature (Stroud and Secrest, 1966) points $\{x_{q_1} : q_1 = 1, \ldots, n_q\}$ and $\{x_{q_2} : q_2 = 1, \ldots, n_q\}$ and weights $\{w_{q_1} : q_1 = 1, \ldots, n_q\}$ and $\{w_{q_2} : q_2 = 1, \ldots, n_q\}$ in terms of standard uniform.

2. Numerically evaluate the joint pmf in (4.8) via the following approximation in a double sum:

$$\sum_{q_1=1}^{n_q}\sum_{q_2=1}^{n_q} w_{q_1}w_{q_2}\prod_{j=1}^{d} f_{X_2j|X_1}(x_{q_2},y_j|x_{q_1})\times \prod_{[jk]\in\mathcal{E}}\frac{f_{jk|X_1X_2}(y_j,y_k|x_{q_1},x_{q_2})}{f_{X_2j|X_1}(x_{q_2},y_j|x_{q_1})f_{X_2k|X_1}(x_{q_2},y_k|x_{q_1})}.$$

Our comparisons show that $n_q = 15$ quadrature points provide good precision for both the 1-factor and 2-factor tree copula models.

4.3 Model selection

In this section we will discuss model selection strategies for the factor tree copula models. Section 4.3.1 proposes vine tree structure selection methods for the residual part of the model that assume the factor tree copula models are constructed with bivariate normal (BVN) copulas. Section 4.3.2 proposes a heuristic algorithm that sequentially selects suitable bivariate copulas to account for any tail dependence/asymmetry as in Kadhem and Nikoloulopoulos (2021a,b).

4.3.1 1-truncated vine tree structure selection

We propose two selection algorithms to choose the 1-truncated vine tree structure $\mathcal{E}$ for the residual part of the model, namely the polychoric and partial selection algorithms. Before that, we provide the necessary tools to form the aforementioned algorithms. These are the estimated polychoric correlations (Olsson, 1979), correlations between each of the items and the first factor and partial correlations between each of the items and the second factor given the first factor (Nikoloulopoulos and Joe, 2015).

The sample polychoric correlation for all possible pairs of items can be estimated as

$$\hat{\rho}_{jk} = \text{argmax}_{\rho} \sum_{i=1}^{n} \log\Big(\Phi_2(\alpha_{j,y_{ij}+1}, \alpha_{k,y_{ik}+1}; \rho) - \Phi_2(\alpha_{j,y_{ij}+1}, \alpha_{k,y_{ik}}; \rho) - \Phi_2(\alpha_{j,y_{ij}}, \alpha_{k,y_{ik}+1}; \rho) + \Phi_2(\alpha_{j,y_{ij}}, \alpha_{k,y_{ik}}; \rho) \Big), \quad 1 \leq j < k \leq d,$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the BVN cdf with correlation parameter ρ.

When all the bivariate copulas are BVN the p-factor copula model is the same as the discretized MVN model with a p-factor correlation matrix, also known as the p-dimensional normal ogive model (Jöreskog and Moustaki, 2001). The 1-factor copula model in (4.3) is the same as the variant of Samejima's (1969) graded response IRT model, known as normal ogive model (McDonald, 1997) with a 1-factor correlation matrix $R = (r_{jk})$ with $r_{jk} = \theta_{1j}\theta_{1k}$ for $j \neq k$. The 2-factor model in (4.5) is the same as the bidimensional (2-factor) normal ogive model with a 2-factor correlation matrix $R = (r_{jk})$ with $r_{jk} = \theta_{1j}\theta_{1k} + \theta_{2j}\theta_{2k}[(1 - \theta_{1j}^2)(1 - \theta_{1k}^2)]^{1/2}$ for $j \neq k$. The parameter θ_{1j} of $C_{X_1 j}$ is the correlation of the underlying normal variable Z_j of Y_j with $Z_{01} = \Phi^{-1}(X_1)$, and the parameter θ_{2j} of $C_{X_2 j}$ is the partial correlation between Z_j and $Z_{02} = \Phi^{-1}(X_1)$ given Z_{01}.

Subsequently, for all possible pair of items we can estimate the partial correlations between Z_j and Z_k given Z_{01} and the partial correlations between Z_j and Z_k given Z_{01}, Z_{02} via the relations

$$\hat{\rho}_{jk;Z_{01}} = \frac{\hat{\rho}_{jk} - \hat{\theta}_{1j}\hat{\theta}_{1k}}{\sqrt{(1 - \hat{\theta}_{1j}^2)(1 - \hat{\theta}_{1k}^2)}} \quad \text{and} \quad \hat{\rho}_{jk;Z_{01},Z_{02}} = \frac{\hat{\rho}_{jk;Z_{01}} - \hat{\theta}_{2j}\hat{\theta}_{2k}}{\sqrt{(1 - \hat{\theta}_{2j}^2)(1 - \hat{\theta}_{2k}^2)}},$$

respectively, where $\hat{\theta}_{1j}, \hat{\theta}_{1k}$ are the estimated unidimensional normal ogive model's parameters and $\hat{\theta}_{1j}, \hat{\theta}_{1k}, \hat{\theta}_{2j}, \hat{\theta}_{2k}$ are the estimated bidimensional normal ogive model's parameters.

The polychoric and partial correlation algorithms select the best vine tree using the minimum spanning tree algorithm (Prim, 1957). The former algorithm selects the edges $\mathcal{E}$ of the tree that minimize the sum of the weights $\log(1-\hat{\rho}^2_{jk})$, while the latter algorithm the sum of the weights $\log(1-\hat{\rho}^2_{jk;Z_{01}})$ for the 1-factor tree copula model and $\log(1-\hat{\rho}^2_{jk;Z_{01},Z_{02}})$ for the 2-factor tree copula model.

4.3.2 Bivariate copula selection

We propose a heuristic method that selects appropriate bivariate copulas for the proposed models. It starts with an initial assumption that all bivariate copulas are BVN and independent copulas in the factor and 1-truncated vine copula model, respectively. Then sequentially suitable copulas with lower or upper tail dependence are assigned where necessary to account for more probability in one or both joint tails. For ease of interpretation, we do not mix Gumbel, s.Gumbel, t_ν and BVN for a single tree of the model; e.g., for the 2-factor tree copula model we allow three different copula families, one for the first factor, one for the second factor and one for the 1-truncated vine (residual dependence part of the model).

The selection algorithm involves the following steps:

1. Start with a factor tree copula model with BVN and independent copulas in the factor and 1-truncated vine copula parts of the model, respectively.

2. Factor part

   (a) Factor 1

i. Fit all the possible models, iterating over all the bivariate copula candidates that link each of the items to X_1.

ii. Select the bivariate copula that corresponds to the highest log-likelihood.

iii. Replace the BVN with the selected bivariate copula that links each of the items to X_1.

(b) Factor 2

i. Fit all the possible models, iterating over all the copula candidates that link each of the items to X_2.

ii. Select the bivariate copula that corresponds to the highest log-likelihood.

iii. Replace BVN with the selected bivariate copula that links each of the items to X_2.

3. 1-truncated vine part

(a) Select the best 1-truncated vine tree structure $\mathcal{E}$ using both the polychoric and partial selection algorithms proposed in Subsection 4.3.1.

(b) Fit all the possible models, iterating over all the bivariate copula candidates that link the pairs of items $\in \mathcal{E}$ given the factors.

(c) Select the bivariate copula that corresponds to the highest log-likelihood.

(d) Replace the independence copula with the selected bivariate copula that links each pair of items $\in \mathcal{E}$ given the factors.

4.4 Simulations

An extensive simulation study is conducted to assess the (a) efficiency of the proposed estimation method and (b) reliability of using the model selection algorithms to select the correct 1-truncated vine tree structure for the residual dependence part of the model. We randomly generated $1,000$ datasets with sample size $n = 500$ and $d = \{8, 16, 24\}$ items with $K = 5$ equally weighted categories from an 1-factor and 2-factor tree copula models with Gumbel copulas. The items in the last tree are either serially connected in ascending order with an 1-truncated drawable vine or randomly connected with a 1-truncated regular vine. Note in passing that the drawable vine is a boundary regular vine case.

We set the copula parameters in Kendall's τ scale, i.e., $\tau(\theta_{1j},\, j = 1, \ldots, d) = \{0.70, \ldots, 0.40\}$ and $\tau(\theta_{2j},\ j = 1, \ldots, d) = \{0.55, \ldots, 0.25\}$ for the factor copula parts of the models and $\tau(\delta_{jk},\, jk \in \mathcal{E}) = \{0.55, \ldots, 0.25\}$ and $\tau(\delta_{jk},\, jk \in \mathcal{E}) = \{0.40, \ldots, 0.10\}$ for the 1-truncated vine copula part of the model for the 1-factor and 2-factor tree copula model, respectively. The τ's as above form equally spaced sequences and are strictly increasing functions of the true (simulated) Gumbel copula parameters, viz.

$$\tau(\theta) = 1 - \theta^{-1}. \tag{4.10}$$

Table 4.1 and Table 4.2 present the resulting biases, standard deviations (SD) and root mean square errors (RMSE), scaled by n, from the simulations of the 1-factor and 2-factor tree copula models with Gumbel copulas, respectively and an 1-truncated drawable vine residual dependence structure. The results indicate that

the proposed approximation method is efficient for estimating the factor tree copula models and the efficiency improves as the dimension increases.

In Figure 4.3 we report the frequency of a pair of items is correctly selected as an edge for each of the edges of the 1-truncated vine from the simulations of the 1- and 2-factor tree copula models with Gumbel copulas with $d = 8$, $d = 16$ and $d = 24$ items for both the partial and polychoric selection algorithms. It has been shown that the partial selection algorithm as the dimension increases performs extremely well for the 1-truncated drawable vine residual dependence structure, but poorly for the 1-truncated regular vine structure. The quite contrary (or complimentary) results are seen for the polychoric algorithm. The polychoric selection algorithm rather performs extremely well in selecting the true edges in the 1-truncated regular vine residual dependence structure. It is most accurate for the initial edges, while it is less accurate for the final edges. This is because the dependence strength is represented in descending order as $\tau = \{0.40, \ldots, 0.10\}$, so the polychoric selection algorithm is highly reliable to select the edges with stronger dependence. The edges with weaker dependence are not easily quantified and can be approximated with other edges that lead to a similar correlation matrix or even accounted for by the previous trees (factor copula models).

Table 4.1: Small sample of size $n = 500$ simulations (10^3 replications) and $d = \{8, 16, 24\}$ items with $K = 5$ equally weighted categories from an 1-factor tree copula model with Gumbel copulas and an 1-truncated drawable vine residual dependence structure for $d = \{8, 16, 24\}$ and resultant biases, root mean square errors (RMSE), and standard deviations (SD), scaled by n, for the IFM estimates.

$d = 8$	1st tree (1-factor copula)								2nd tree (vine copula)						
τ	0.70	0.66	0.61	0.57	0.53	0.49	0.44	0.40	0.40	0.35	0.30	0.25	0.20	0.15	0.10
nBias	6.19	5.83	8.34	7.30	4.13	-0.46	-2.47	-2.77	-14.23	-16.11	-15.79	-9.90	-2.86	1.19	1.42
nSD	20.48	21.24	19.05	17.56	16.43	16.56	15.79	16.05	44.97	33.61	28.66	25.17	21.68	19.87	18.54
nRMSE	21.40	22.03	20.80	19.01	16.94	16.57	15.98	16.29	47.17	37.27	32.72	27.05	21.87	19.91	18.60

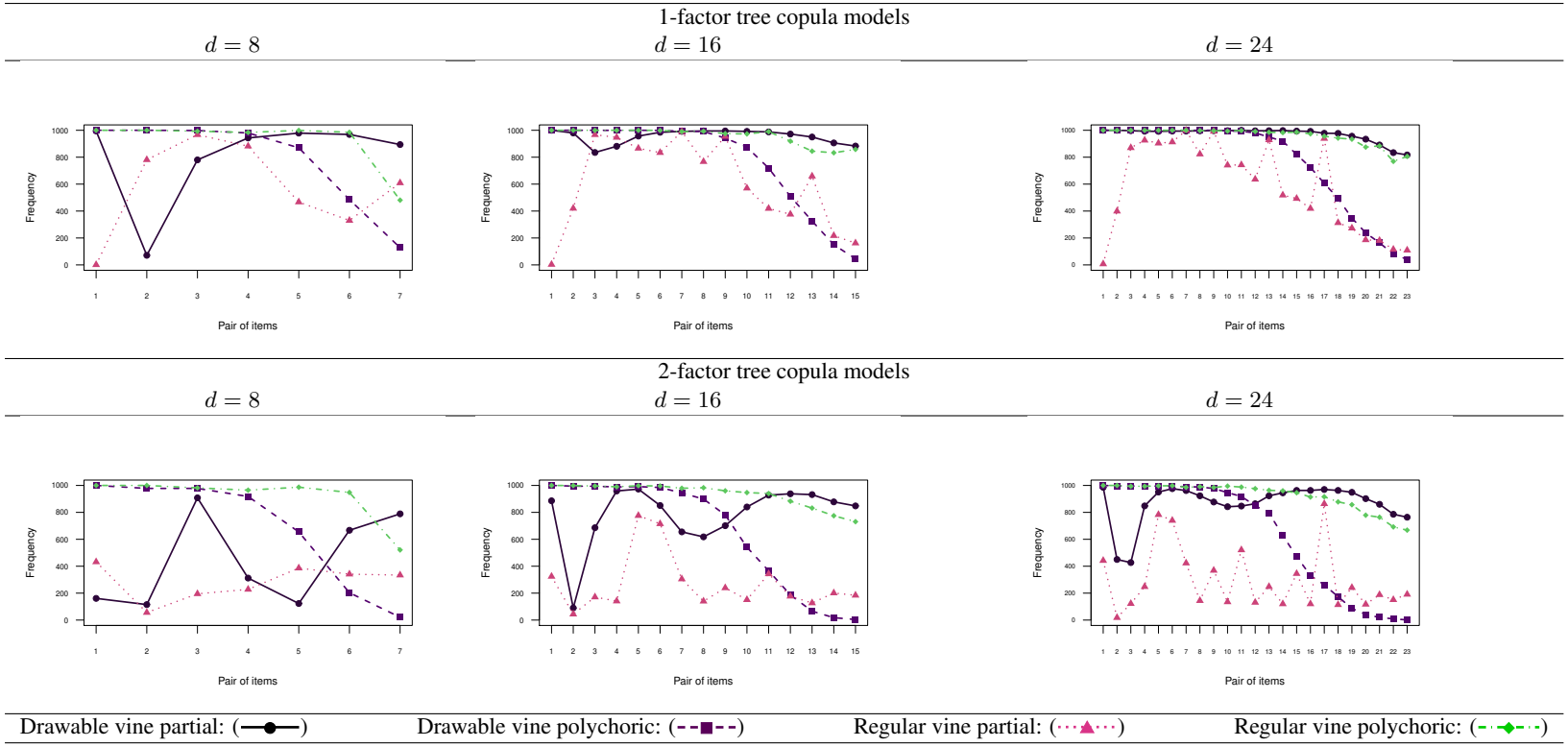
$d = 16$	1st tree (1-factor copula)														
τ	0.70	0.68	0.66	0.64	0.62	0.60	0.58	0.56	0.54	0.52	0.50	0.48	0.46	0.44	0.42	0.40
nBias	2.76	3.43	5.22	6.18	6.02	4.66	2.96	2.19	0.79	0.20	0.05	-1.43	-1.74	-1.02	-1.80	-0.93
nSD	10.89	11.31	11.85	11.94	12.08	11.91	12.35	12.45	12.65	13.26	12.96	13.66	13.66	14.51	14.55	14.19
nRMSE	11.23	11.81	12.95	13.45	13.49	12.79	12.70	12.64	12.68	13.26	12.96	13.74	13.77	14.55	14.66	14.22
	2nd tree (1-truncated vine copula)														
τ	0.40	0.38	0.36	0.34	0.31	0.29	0.27	0.25	0.23	0.21	0.19	0.16	0.14	0.12	0.10	
nBias	-6.55	-9.58	-12.27	-11.32	-9.85	-6.42	-4.51	-2.46	-1.01	0.46	0.70	1.35	1.96	1.17	1.59	
nSD	22.62	22.71	21.92	20.66	19.36	18.59	18.95	18.22	17.92	18.02	17.21	17.20	16.79	16.91	16.62	
nRMSE	23.55	24.65	25.12	23.56	21.72	19.67	19.48	18.39	17.95	18.02	17.22	17.25	16.90	16.95	16.70	

$d = 24$	1st tree (1-factor copula)																							
τ	0.70	0.69	0.67	0.66	0.65	0.63	0.62	0.61	0.60	0.58	0.57	0.56	0.54	0.53	0.52	0.50	0.49	0.48	0.47	0.45	0.44	0.43	0.41	0.40
nBias	1.61	1.89	3.41	4.20	4.35	3.84	3.13	2.52	2.29	1.68	1.03	0.44	-0.21	0.05	-0.53	-0.55	-0.28	-0.05	-0.12	-0.33	-0.44	-0.12	-0.25	-0.60
nSD	9.72	10.39	10.86	11.06	11.13	10.86	11.28	11.32	11.61	11.99	11.76	11.90	12.10	12.54	12.71	12.70	12.82	13.21	13.54	13.43	13.86	13.74	13.57	13.84
nRMSE	9.86	10.56	11.38	11.83	11.95	11.52	11.70	11.59	11.83	12.11	11.80	11.91	12.11	12.54	12.73	12.71	12.82	13.21	13.54	13.43	13.87	13.74	13.58	13.85
	2nd tree (1-truncated vine copula)																							
τ	0.40	0.39	0.37	0.36	0.35	0.33	0.32	0.30	0.29	0.28	0.26	0.25	0.24	0.22	0.21	0.20	0.18	0.17	0.15	0.14	0.13	0.11	0.10	
nBias	-4.29	-6.22	-7.94	-8.53	-7.72	-6.13	-6.24	-4.19	-2.61	-2.03	-1.33	-0.34	0.17	-0.32	0.59	-0.06	0.44	1.32	0.74	0.60	0.46	0.04	0.73	
nSD	20.39	19.93	19.75	19.11	19.40	18.36	18.93	18.26	18.15	18.02	17.04	17.40	16.77	17.03	17.64	16.52	17.28	16.69	17.22	16.72	17.12	16.88	16.79	
nRMSE	20.84	20.88	21.28	20.93	20.88	19.36	19.94	18.73	18.33	18.14	17.10	17.41	16.78	17.04	17.65	16.52	17.29	16.74	17.24	16.73	17.13	16.88	16.80	

Table 4.2: Small sample of size $n = 500$ simulations (10^3 replications) and $d = 24$ items with $K = 5$ equally weighted categories from a 2-factor tree copula model with Gumbel copulas and an 1-truncated drawable vine residual dependence structure and resultant biases, root mean square errors (RMSE), and standard deviations (SD), scaled by n, for the IFM estimates.

$d = 24$																								
	1st tree (1st factor of 2-factor copula)																							
τ	0.70	0.69	0.67	0.66	0.65	0.63	0.62	0.61	0.60	0.58	0.57	0.56	0.54	0.53	0.52	0.50	0.49	0.48	0.47	0.45	0.44	0.43	0.41	0.40
nBias	-5.74	-3.26	-0.07	2.35	3.96	4.12	3.60	3.94	4.05	3.73	4.58	4.27	3.74	4.83	4.17	5.08	4.28	4.56	5.15	4.80	4.82	4.05	4.42	2.96
nSD	26.55	26.96	27.90	27.43	25.80	24.89	25.15	24.57	23.62	23.93	23.89	23.53	23.21	23.04	22.38	23.15	22.39	23.75	22.93	22.04	22.38	21.99	22.71	21.74
nRMSE	27.16	27.15	27.90	27.53	26.11	25.23	25.41	24.89	23.97	24.22	24.33	23.91	23.51	23.54	22.77	23.70	22.80	24.18	23.50	22.56	22.89	22.36	23.14	21.94
	2nd tree (2nd factor of 2-factor copula)																							
τ	0.55	0.54	0.52	0.51	0.50	0.48	0.47	0.46	0.45	0.43	0.42	0.41	0.39	0.38	0.37	0.35	0.34	0.33	0.32	0.30	0.29	0.28	0.26	0.25
nBias	4.31	1.24	2.81	0.39	-0.58	-1.81	-2.58	-3.06	-6.03	-6.58	-8.23	-9.13	-9.58	-12.73	-13.14	-11.90	-9.67	-10.48	-12.89	-11.57	-11.57	-12.77	-11.14	-8.04
nSD	40.65	41.80	42.93	45.05	43.16	42.69	41.67	40.68	40.38	41.00	41.35	39.73	41.24	41.35	40.48	40.60	41.84	42.41	40.90	38.62	40.15	37.78	39.96	38.41
nRMSE	40.88	41.82	43.02	45.05	43.17	42.73	41.75	40.79	40.83	41.52	42.16	40.76	42.34	43.27	42.56	42.31	42.94	43.68	42.88	40.31	41.78	39.88	41.49	39.25
	3rd tree (1-truncated vine copula)																							
τ	0.40	0.39	0.37	0.36	0.35	0.33	0.32	0.30	0.29	0.28	0.26	0.25	0.24	0.22	0.21	0.20	0.18	0.17	0.15	0.14	0.13	0.11	0.10	
nBias	0.10	-4.49	-9.56	-10.74	-9.52	-9.21	-6.47	-4.90	-2.94	-3.25	-0.50	-0.21	0.85	1.52	2.04	0.34	1.66	1.66	1.76	2.45	2.02	2.29	2.25	
nSD	32.64	35.17	31.46	28.61	27.74	24.35	24.49	22.53	25.08	23.54	22.79	20.38	21.06	20.56	20.37	22.01	20.16	20.08	19.14	19.56	18.21	18.11	18.33	
nRMSE	32.64	35.46	32.88	30.56	29.33	26.03	25.33	23.06	25.25	23.76	22.80	20.38	21.07	20.61	20.48	22.01	20.23	20.15	19.22	19.71	18.33	18.25	18.47	

Figure 4.3: Small sample of size $n = 500$ simulations (10^3 replications) and $d = \{8, 16, 24\}$ items with $K = 5$ equally weighted categories from 1-factor and 2-factor tree copula models with Gumbel copulas and an 1-truncated drawable/regular vine residual dependence structure and resultant number of times a pair of items is correctly selected as an edge for each of the edges of the 1-truncated drawable and regular vine copula for both the partial and polychoric selection algorithms.



4.5 Application

In this section we illustrate the proposed methodology by analysing $d = 20$ items from a subsample of $n = 221$ veterans who reported clinically significant Post Traumatic Stress Disorder (PTSD) symptoms (Armour et al., 2017). PTSD can be defined as a mental disorder associated with extreme distress and disruption of daily activities as a result of experiencing or witnessing a traumatic event. The PTSD items are divided into four domains: (1) intrusions (e.g., repeated, disturbing and unwanted memories), (2) avoidance (e.g., avoiding external reminder of the stressful experience), (3) cognition and mood alterations (e.g., trouble remembering important parts of the stressful experience) and (4) reactivity alterations (e.g., taking too many risks or doing things that could cause you harm). Each item is answered in a five-point ordinal scale: "0 = Not at all", "1 = A little bit", "2 = Moderately", "3 = Quite a bit" and "4 = Extremely". The dataset and its complete description can be found in Table 4.3 and Armour et al. (2017) or in the R package **BGGM** (Williams and Mulder, 2020).

For some items, it is plausible that a veteran might be thinking about the maximum trauma (or a high quantile) of many past events. For example, for the items in the first domain, a participant might reflect on past relevant events where an intrusion affected their life; then by considering the worst case, i.e., the event where the negative effect of an intrusion in their life was substantial, they choose an appropriate ordinal response. For some of the other items, one might consider a median or less extreme harm of past relevant events. To sum up, the items appear to be a mixed selection between discretized averages and maxima so that a factor model with more

Table 4.3: The Post Traumatic Stress Disorder (PTSD) with 20 items categorized into 4 groups.

Number	Item	Group
1	Intrusive Thoughts	Intrusions
2	Nightmares	
3	Flashbacks	
4	Emotional cue reactivity	
5	Psychological cue reactivity	
6	Avoidance of thoughts	Avoidance
7	Avoidance of reminders	
8	Trauma-related amnesia	Cognition and mood alterations
9	Negative beliefs	
10	Blame of self or others	
11	Negative trauma-related emotions	
12	Loss of interest	
13	Detachment	
14	Restricted affect	
15	Irritability/anger	Arousal and reactivity alterations
16	Self-destructive/reckless behavior	
17	Hypervigilance	
18	Exaggerated startle response	
19	Difficulty concentrating	
20	Sleep disturbance	

probability in the joint upper tail might be an improvement over a factor model based on a discretized MVN.

The interpretations as above suggest that a factor tree with a combination of Gumbel and BVN or t_ν copulas might provide a better fit. To further explore the above interpretations, we calculate the average of lower and upper polychoric semi-correlations (Kadhem and Nikoloulopoulos, 2021a,b) for all variables to check if there is any overall tail asymmetry. For comparison, we also report the theoretical semi-correlations under different choices of copulas. Table 4.4 shows averages of the polychoric semi-correlations for all pairs along with the theoretical semi-correlations under different choices of copulas. Overall, we see that there is more correlation in

the joint upper tail than the joint lower tail, suggesting that factor tree copula models with Gumbel bivariate copulas might be plausible.

Table 4.4: Average observed polychoric correlations and semi-correlations for all pairs of items for the Post Traumatic Stress Disorder dataset, along with the corresponding theoretical semi-correlations for BVN, t_2, t_5, Frank, Gumbel , and survival Gumbel (s.Gumbel) copulas.

	ρ_N	ρ_N^-	ρ_N^+
Observed	0.35	0.26	0.47
BVN	0.35	0.16	0.16
t_2	0.35	0.49	0.49
t_5	0.35	0.35	0.35
Frank	0.35	0.10	0.10
Gumbel	0.35	0.11	0.37
s.Gumbel	0.35	0.37	0.11

We then select a suitable vine tree structure using the partial and polychoric selection algorithms proposed in Section 4.3.1 and compute various discrepancy measures between the observed polychoric correlation matrix $\mathbf{R}_{\text{observed}}$ and the correlation matrix $\mathbf{R}_{\text{model}}$ based on factor tree copula models with BVN copulas. We report the maximum absolute correlation difference $D_1 = \max|\mathbf{R}_{\text{model}} - \mathbf{R}_{\text{observed}}|$, the average absolute correlation difference $D_2 = \text{avg}|\mathbf{R}_{\text{model}} - \mathbf{R}_{\text{observed}}|$ and the correlation matrix discrepancy measure $D_3 = \log\big(\det(\mathbf{R}_{\text{model}})\big) - \log\big(\det(\mathbf{R}_{\text{observed}})\big) + \text{tr}(\mathbf{R}_{\text{model}}^{-1}\mathbf{R}_{\text{observed}}) - d$. For a baseline comparison, we also compute the discrepancy measures for the 1- and 2-factor copula models with BVN copulas. We aim to obtain a dependence structure that results in the lowest discrepancy measure; this will indicate a suitable vine structure for the item response data on hand.

After finding a suitable vine structure, we construct a plausible factor tree copula model, to analyse any type of items, by using the proposed heuristic algorithm in Section 4.3.2. We use the AIC at the IFM estimates as a rough diagnostic measure for

model selection between the models. In addition, we use the Vuong (1989) procedure that is based on the sample version of the difference in Kullback-Leibler divergence.

Note in passing that the 2-factor (tree) copula models with BVN copulas will have one dependence parameter less as one copula in the second factor is set to independence for identification purposes.

Table 4.5 shows that the sample correlation matrix of the data has a 2-factor tree structure according to the discrepancy measures. The table also gives the AICs and the 95% CIs of Vuong's tests for all the fitted models. The best fitted model, based on AIC values, is the 2-factor tree copula model obtained from the partial selection algorithm. The best fitted 2-factor tree copula model has the t_2 for the 1st tree, Gumbel for the 2nd tree, and t_5 for the 3rd tree. From the Vuong's 95% Cls it is shown that 2-factor tree copula model provides a big improvement over its Gaussian analogue and outperforms all the other fitted models except the 2-factor tree obtained from the polychoric selection algorithm. The tree selection algorithms might not yield into the same 'true' vine tree, however, closely approximated factor tree copula models are achieved. The factor tree copula model is mostly constructed with t_2 bivariate copulas which are suitable for both positive and negative dependence, however the highest dependence is found in the 2nd factor which is constructed with Gumbel copulas. This is in line with both the initial interpretations and preliminary analysis which suggest that some items can be considered as discretized maxima.

Table 4.6 includes the copula parameter estimates in Kendall's τ scale and their standard errors (SE) for the selected 2-factor and 2-factor tree copula models. The latter is obtained from the partial selection algorithm. To make it easier to compare

Table 4.5: Measures of discrepancy between the sample and the resulting correlation matrix from the 1-factor, 2-factor, 1-factor tree, and 2-factor tree copula models with BVN copulas for the Post Traumatic Stress Disorder dataset, along with the AICs, Vuong's 95% CIs, for the 1-factor, 2-factor, 1-factor tree, and 2-factor tree copula models with BVN and selected copulas. Alg.1: partial selection algorithm; Alg.2: polychoric selection algorithm.

	Factor copula		1-factor tree copula		2-factor tree copula	
	1-factor	2-factor	Alg.1	Alg.2	Alg.1	Alg.2
BVN copulas						
D_1	0.40	0.30	0.23	0.20	0.15	0.20
D_2	0.08	0.05	0.05	0.05	0.03	0.05
D_3	4.53	2.80	1.75	1.83	1.17	1.75
#parameters	20	39	39	39	58	58
AIC	12031.1	11764.0	11632.4	11642.1	11549.1	11611.8
Selected copulas						
#parameters	20	40	39	39	59	59
AIC	11800.4	11413.5	11355.3	11344.89	11189.1	11240.3
Vuong's 95% CI[a]	(0.21, 0.63)	(0.25, 0.79)	(0.37, 0.89)	(0.43, 0.91)	(0.54, 1.09)	(0.58, 1.11)
Vuong's 95% CI[b]	(1.50, 2.31)	(0.99, 1.67)	(0.79, 1.40)	(0.83, 1.40)	-	(0.69, 1.24)
Vuong's 95% CI[c]	(1.17, 1.80)	(0.60, 1.02)	(0.30, 0.63)	(0.27, 0.61)	-	(-0.002, 0.23)

[a] Selected factor (tree) copula models versus their Gaussian analogues.
[b] Selected 2-factor tree copula model with Alg.1 versus other fitted models with BVN copulas.
[c] Selected 2-factor tree copula model with Alg.1 versus other fitted models with selected copulas.

Table 4.6: Estimated copula parameters and their standard errors (SE) in Kendall's τ scale for the selected 2-factor and 2-factor tree copula models obtained from the partial selection algorithm for the Post Traumatic Stress Disorder dataset.

	2-factor copula				2-factor tree copula						
Tree	1st factor		2nd factor		1st factor		2nd factor		Vine model		
Copula	t_2		Gumbel		t_2		Gumbel		t_5		
Items	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\hat{\tau}$	SE	$\mathcal{E}$	$\hat{\tau}$	SE
1	0.16	0.06	0.49	0.04	-0.17	0.06	0.50	0.04	$1, 18$	-0.18	0.06
2	0.11	0.06	0.49	0.04	-0.08	0.06	0.45	0.04	$18, 17$	0.22	0.06
3	0.14	0.06	0.54	0.04	-0.12	0.06	0.52	0.04	$18, 14$	-0.20	0.07
4	0.32	0.06	0.56	0.05	-0.34	0.06	0.57	0.05	$18, 10$	-0.10	0.06
5	0.21	0.06	0.55	0.04	-0.21	0.06	0.56	0.04	$10, 11$	0.36	0.05
6	0.13	0.06	0.28	0.05	-0.13	0.06	0.26	0.05	$11, 9$	0.29	0.06
7	0.11	0.06	0.40	0.04	-0.09	0.06	0.39	0.04	$9, 2$	-0.18	0.06
8	-0.03	0.06	0.21	0.05	0.04	0.06	0.19	0.05	$2, 3$	0.26	0.06
9	-0.17	0.06	0.38	0.04	0.24	0.06	0.33	0.04	$3, 20$	0.05	0.07
10	0.16	0.06	0.34	0.05	-0.12	0.06	0.30	0.04	$2, 16$	0.13	0.06
11	0.09	0.06	0.52	0.04	-0.07	0.06	0.48	0.04	$16, 15$	0.17	0.06
12	-0.23	0.06	0.50	0.04	0.28	0.06	0.50	0.04	$9, 4$	0.29	0.08
13	-0.35	0.06	0.55	0.05	0.34	0.05	0.49	0.05	$20, 5$	0.05	0.07
14	-0.37	0.05	0.41	0.05	0.35	0.05	0.36	0.05	$14, 13$	0.27	0.07
15	-0.09	0.06	0.48	0.04	0.11	0.06	0.44	0.04	$5, 6$	0.12	0.07
16	-0.08	0.06	0.31	0.05	0.10	0.06	0.28	0.04	$6, 7$	0.23	0.06
17	-0.04	0.06	0.34	0.04	0.04	0.06	0.33	0.04	$7, 19$	-0.21	0.06
18	-0.06	0.06	0.45	0.04	0.12	0.06	0.46	0.04	$16, 8$	0.12	0.06
19	-0.26	0.06	0.45	0.04	0.28	0.06	0.43	0.04	$19, 12$	0.08	0.07
20	-0.11	0.06	0.41	0.04	0.13	0.06	0.40	0.04	-	-	-

strengths of dependence, we convert the BVN/t_ν and Gumbel/s.Gumbel copula parameters to Kendall's τ's via the relation $\tau(\theta) = \frac{2}{\pi}\arcsin(\theta)$ and (4.10), respectively. Interestingly, the Kendall's τ's in the 2-factor copula model are roughly equivalent to the estimates in the 1st and 2nd factors of the 2-factor tree copula model. Most of the dependence is captured in the first two trees, resulting in weak to medium residual dependencies in the 1-truncated vine copula model, but significantly larger from independence. Overall, the items, in the Markov tree, are mostly positively associated to one another with only few negative conditional dependencies. The residual dependencies reveal that there is stronger association between the 10th and 11th items

that are “Blame of self or others” and “Negative trauma-related emotions”, respectively. In addition, there is moderate association between items 9 and 11 that are “Negative beliefs” and “Negative trauma-related emotions”, respectively. With similar moderate dependence found between items 9 and 4 that are “Negative beliefs” and “Emotional cue reactivity”, respectively.

4.6 Chapter summary

In this chapter, we have proposed factor tree copula models for item response data. These are truncated vine copula models that involve both observed and latent variables. This construction allows for conditional dependence of observed variables given very few latent variables.

The proposed models preserve the flexible dependence properties of the factor/vine copulas. They are parsimonious models and offer dependence modelling with different tail behaviour. We consider the proposed combined factor/truncated vine structure to be reasonable as most of the dependence is explained via a factor copula model and any residual dependence is captured by an additional layer of dependence in the form of an 1-truncated vine copula.

Chapter 5

Discussion and future research

Factor copula models can provide flexible reflection asymmetric tail and non-linear dependence. They are parsimonious models and favourable for large dimensions, so the number of parameters is $\mathcal{O}(d)$ instead of $\mathcal{O}(d^2)$. Factor copulas can be viewed as a truncated canonical vine copulas (Brechmann et al., 2012) rooted at the latent variables, that are constructed from a sequence of bivariate copulas in hierarchies or tree levels. Joe et al. (2010) show that in order for a vine copula to have (tail) dependence for all bivariate margins, it is only necessary for the bivariate copulas in level 1 to have (tail) dependence and not necessary for the conditional bivariate copulas in levels $2, \ldots, d-1$ to have tail dependence.

In this thesis, we made new contributions in proposing several extensions of factor copula models, along with model selection algorithms and goodness-of-fit statistics. For clarity, the proposed modelling frameworks are summarised and discussed in the subsequent sections along with a preview of future research, followed with final remarks.

5.1 Factor copula models for mixed data

In Chapter 2, we have extended the factor copula model proposed in Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015) to the case of mixed continuous and discrete responses. It is the most general factor model as (a) it has the standard factor model with an additive latent structure as a special case when the BVN copulas are used, (b) it can have a latent structure that is not additive if other than BVN copulas are called, (c) the parameters of the univariate distributions are separated from the copula (dependence) parameters which are interpretable as dependence of an observed variable with a latent variable, or conditional dependence of an observed variable with a latent variable given preceding latent variables. Other non-linear (e.g., Rizopoulos and Moustaki 2008), semi- (e.g., Gruhl et al. 2013) or non-parametric models (e.g., Kelava et al. 2017) with latent variables have either an additive latent structure or allow polynomial and interaction terms to be added in the linear predictor, hence are not as general. Another mixed-variable model in the literature that is called a factor copula model (Murray et al., 2013) is restricted to the MVN copula as the model proposed by Gruhl et al. (2013), hence has an additive latent structure.

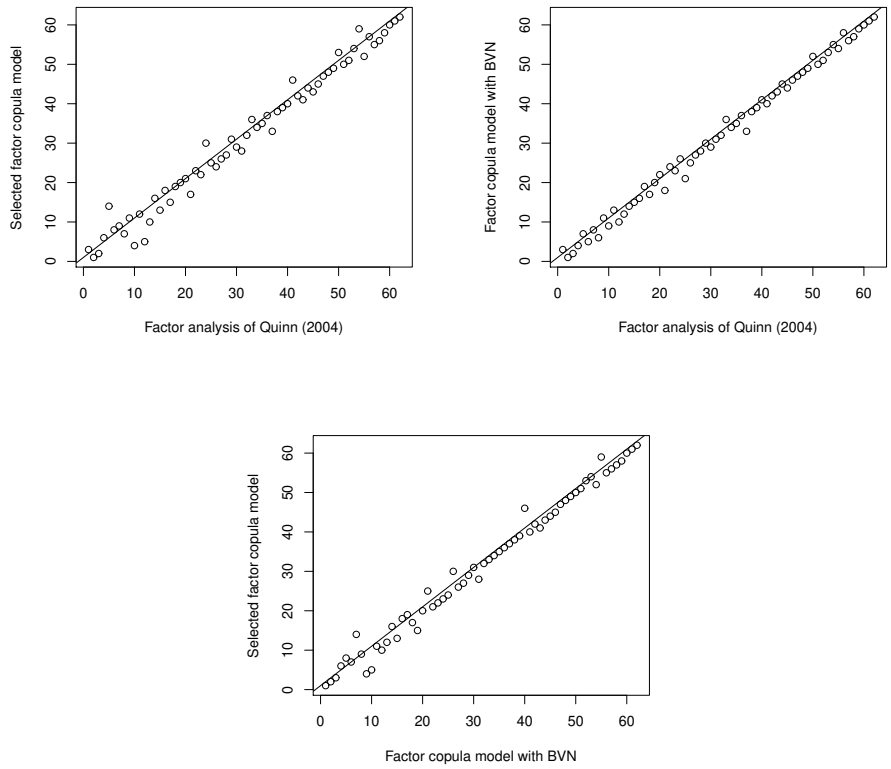
We have shown that factor copula models provide a substantial improvement over the standard factor model on the basis of the log-likelihood principle, Vuong's and M_2 statistics. Hence, superior statistical inference for the loading parameters of interest can be achieved. This improvement relies on the fact that the latent variable distribution is expressed via factor copulas instead of the MVN distribution. The latter is restricted to linear and reflection symmetric dependence. Rizopoulos and Moustaki (2008) stressed that the inadequacy of normally distributed latent variables can be caused by the non-linear dependence on the latent variables. The factor copula

can provide flexible reflection asymmetric tail and non-linear dependence as it is a truncated canonical vine copula (Brechmann et al., 2012) rooted at the latent variables. The 1-factor copula has bivariate copulas with tail dependence in the 1st level and independence copulas in all the remaining levels of the vine (truncated after the 1st level). The 2-factor copula has bivariate copulas with tail dependence in the 1st and 2nd level and independence copulas in all the remaining levels (truncated after the 2nd level). Hence, the tail dependence among the latent variables and each of the observed variables is passed to the tail dependence among the observed variables.

Even in case where the effect of misspecifying the bivariate linking copula to build the factor copula models can be seen as minimal for the Kendall's τ (loading) parameters, the tail dependence varies, as explained in Section 1.3, and is a property to consider when choosing amongt different families of copulas and hence affects prediction. Rabe-Hesketh et al. (2003) highlighted the importance of the correct distributional assumptions for the prediction of latent scores. The latent scores will essentially show the effect of different model assumptions, because it is an inference that depends on the joint distribution. Factor copula models have bivariate copulas that link the latent variables to each of the observed variables. If these bivariate copulas have upper or lower tail dependence, then this type of dependence is passed to the dependence between the factor scores and each of the observed variables. Hence, factor scores are fairly different than the ones for the standard factor model if the sample size is sufficient. Figure 5.1 demonstrates these differences by revisiting the political-economic dataset in Section 2.5.1 and comparing the political-economic risk ranking obtained via our selected model, the factor copula model with BVN copulas (standard factor model), and the mixed-data factor analysis of Quinn

(2004). It is revealed that even for a small sample size ($n = 62$) there are differences. Between the factor copula model with BVN copulas and the factor analysis model of Quinn (2004), there are small to moderate differences, because while these models share the same latent variables distribution, the former model does not assume the observed variables to be normally distributed, but rather uses the empirical distribution of the continuous observed variables, i.e. allows the margins to be quite free and not restricted by normal distribution. The differences in the lower panel graph are solely due the miss-specification the latent variable distribution.

Figure 5.1: Comparison of the political-economic risk rankings obtained via our selected model, the standard factor model, and the mixed-data factor analysis of Quinn (2004).



As stated by many researchers (e.g., Rabe-Hesketh and Skrondal 2001; Skrondal and Rabe-Hesketh 2004), the major difficulty of all the models with latent variables is identifiability. For example, for the standard factor model or the more flexible model in Irincheeva et al. (2012b) one of loadings in the second factor has to be set to zero, because the model with $2d$ loadings is not identifiable. The standard factor model arises as special case of our model if we use as bivariate linking copulas the BVN copulas. Hence, for the 2-factor copula model with BVN copulas, one of the BVN copulas in the second factor has to be set as an independence copula. However, using other than BVN copulas, the 2-factor copula model is near-identifiable with $2d$ bivariate linking copulas as it has been demonstrated by Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015).

5.2 Structured factor copula models for item response data

For item response data that can be split into non-overlapping groups, we have proposed bi-factor and second-order copula models where we replace BVN distributions, between observed and latent variables, with bivariate copulas. Our copula constructions include the Gaussian bi-factor and second-order models as special cases and can provide a substantial improvement over the Gaussian models based on AIC, Vuong's and goodness-of-fit statistics. Hence, superior statistical inference for the loading parameters of interest can be achieved. The improvement relies on the fact that when we use appropriate bivariate copulas other than BVN copulas in the construction, there is an interpretation of latent variables that can be maxima/minima or mixture of means instead of means.

Our constructions have a latent structure that is not additive as in (3.3) and (3.4) if other than BVN copulas are called and the bi-factor copula (dependence) parameters are interpretable as dependence of an observed variable with the common factor, or conditional dependence of an observed variable with the group-specific latent variable given the common factor.

We have proposed a fast and efficient likelihood estimation technique based on Gauss-Legendre quadrature points. The joint pmfs in (3.1) and (3.2) reduce to one-dimensional integrals of a function which in turn is a product of G one-dimensional integrals. Hence, the evaluation of the joint likelihood requires only low-dimensional integration, as in the one- and two-factor copula models, regardless of the dimension $G + 1$ of the factors. This is an advantage over the p-factor ($p > 2$) copula models where the joint pmf requires p-dimensional integration and becomes intractable as the number of factors increases. Hence, the proposed structured multidimensional factor models provide parsimonious factor solutions without any computational deficiencies as in the p-factor copula models when p increases.

Building on the bi-factor and second-order copula models in Chapter 3, there are several extensions that can be implemented. The adoption of the structure of the Gaussian tri-factor and the third-order models (e.g., Rijmen et al. 2014), to account for any additional layer of dependence, is feasible using the notion of truncated vine copulas that involve both observed and latent variables.

5.3 Factor tree copula models for item response data

We have proposed combined factor/truncated vine copula models to capture the residual dependence for item response data. Due to residual dependencies, the factor

copula models might be too parsimonious as they are restricted to the conditional independence assumption. By combining the factor copula models with an 1-truncated vine copula model, we construct conditional dependence models given very few interpretable latent variables. The combined factor/truncated vine structure has the form of (i) primary dependence being explained by one or more latent variables, and (ii) conditional dependence of observed variables given the latent variables (Joe, 2018).

We have shown that the proposed models provide a substantial improvement over the 1-factor and 2-factor copula models with BVN and selected copulas on the basis of the AIC and Vuong's statistics. We consider the 1-factor and 2-factor tree copula models to be reasonable parsimonious models as most of the dependence is explained via the first few trees in the factor model. This is because that for all the bivariate margins to have upper/lower tail dependence, it only suffices that the bivariate copulas in the first trees (factor part) to have upper/lower tail dependence and is not necessary for the bivariate copulas in the higher trees after the 1-truncated vine to have tail dependence (Joe et al., 2010).

The proposed combined factor with 1-truncated vine copula models in Chapter 4 can be extended to variety of parsimonious factor and vine models. For example, the bi-factor or second-order models for non-overlapping groups of items (See e.g., Gibbons and Hedeker 1992; Gibbons et al. 2007; Kadhem and Nikoloulopoulos 2021a) can be combined with vine copula models to capture the residual dependence for item response data in overlapping groups.

5.4 Final remarks

Although the proposed models require bi-dimensional integration, the evaluation of their likelihood might be time consuming for high-dimensional data. Krupskii and Joe (2022) have shown that proxy variables that are unweighted averages computed from the observed variables can be used for the latent variables when the dimension is large. Alternative log-likelihoods without integrals can be used for parameter estimation and the proxy variables can help to select appropriate linking copulas in some factor copula models and to perform numerically faster maximum likelihood estimation of parameters.

Selecting a suitable copula model via minimizing AIC is extensively used in the copula literature (see, e.g., Nagler et al. 2019; Panagiotelis et al. 2017; Joe 2014; Dißmann et al. 2013; Czado et al. 2013). In our work we have followed a similar approach in optimising the AIC for the search of a suitable model in Chapters 2, 3 and 4. Nevertheless, as the model structure is tuned by optimising the AIC there is a risk of over-fitting the AIC (Cawley and Talbot, 2010), which can be viewed a random variable and will vary from one sample of data to another. This might be substantial if many bivariate copula choices are made. Thus, we only use a few bivariate linking copulas that have distinct tail dependence properties. We have also developed simple diagnostics based on semi-correlations to identify a plausible model. In the data examples in this thesis, the conclusions from the simple diagnostics strongly agree with the conclusions from the proposed model selection algorithms. This, in conjunction with the extensive simulated evidence about the performance of the model selection algorithms, shows that the proposed model selection algorithms can be called without any caveat.

Bibliography

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. John Wiley & Sons.

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., and Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, 45:49–59.

Bagby, R., Parker, J. D., and Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale–I. item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, 38(1):23–32.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley Series in Probability and Statistics. Wiley.

Bedford, T. and Cooke, R. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1):245–268.

Bedford, T. and Cooke, R. M. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031 – 1068.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika*, 76(1):57–76.

Braeken, J., Kuppens, P., Boeck, P. D., and Tuerlinckx, F. (2013). Contextualized personality questionnaires: A case for copulas in structural equation models for categorical data. *Multivariate Behavioral Research*, 48(6):845–870.

Braeken, J., Tuerlinckx, F., and De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, 72(3):393–411.

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

Brechmann, E. C. and Joe, H. (2014). Parsimonious parameterization of correlation matrices using truncated vines and factor analysis. *Computational Statistics & Data Analysis*, 77:233–251.

Briganti, G. and Linkowski, P. (2020). Network approach to items and domains from the Toronto Alexithymia Scale. *Psychological Reports*, 123(5):2038–2052.

Cai, B., Lin, X., and Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis*, 55(9):2644–2651.

Cawley, G. C. and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

Chang, B. and Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics & Data Analysis*, 139:45–63.

Chen, W.-H. and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289.

Czado, C., Jeske, S., and Hofmann, M. (2013). Selection strategies for regular vine copulae. *Journal de la société française de statistique*, 154(1):174–191.

de la Torre, J. and Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 33(8):620–639.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43(2):145–168.

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

Fréchet, M. (1951). Sur les tableaux de correlationdont les marges sont données. *Ann.Univ. Lyon Sc.*, 4:53–84.

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

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

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., and Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31(1):4–19.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3):423–436.

Gignac, G. E., Palmer, B. R., and Stough, C. (2007). A confirmatory factor analytic investigation of the TAS–20: Corroboration of a five-factor model and suggestions for improvement. *Journal of Personality Assessment*, 89(3):247–257.

Gruhl, J., Erosheva, E. A., and Crane, P. K. (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *The Annals of Applied Statistics*, 7(4):2361–2383.

Gustafsson, J.-E. and Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, 28(4):407–434. PMID: 26801141.

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

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283.

Hua, L. and Joe, H. (2011). Tail order and intermediate tail dependence of multivariate copulas. *Journal of Multivariate Analysis*, 102(10):1454–1471.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(4):893–908.

Irincheeva, I., Cantoni, E., and Genton, M. (2012a). A non-Gaussian spatial generalized linear latent variable model. *Journal of Agricaltural, Biological and Environmetal statistics*, 17:332–353.

Irincheeva, I., Cantoni, E., and Genton, M. G. (2012b). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, 39(4):663–680.

Jiryaie, F., Withanage, N., Wu, B., and de Leon, A. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation*, 86(9):1643–1659.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, 46(2):262 – 282.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In Rüschendorf, L., Schweizer, B., and Taylor, M. D., editors, *Distributions with Fixed Marginals and Related Topics*,

volume 28, pages 120–141, Hayward, CA. Institute of Mathematical Statistics, Institute of Mathematical Statistics.

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

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Joe, H. (2011). Tail dependence in vine copulae. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*, pages 165–187, Singapore. World Scientific.

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman and Hall/CRC.

Joe, H. (2018). Parsimonious graphical dependence models constructed from vines. *Canadian Journal of Statistics*, 46(4):532–555.

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

Jöreskog, K. G. and Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research*, 36:347–387.

Kadhem, S. H. and Nikoloulopoulos, A. K. (2021a). Bi-factor and second-order copula models for item response data. *ArXiv e-prints*, arXiv:2102.10660.

Kadhem, S. H. and Nikoloulopoulos, A. K. (2021b). Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, 74(3):365–403.

Kadhem, S. H. and Nikoloulopoulos, A. K. (2021c). *FactorCopula: Factor, Bi-Factor and Second-Order Copula Models*. R package version 0.8. URL: http://CRAN.R-project.org/package=FactorCopula.

Kadhem, S. H. and Nikoloulopoulos, A. K. (2022). Factor tree copula models for item response data. *ArXiv e-prints*, arXiv:2201.00339.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, 56(12):4243–4258.

Kelava, A., Kohler, M., Krzyźak, A., and Schaffland, T. F. (2017). Nonparametric estimation of a latent variable model. *Journal of Multivariate Analysis*, 154:112–134.

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6):2836–2850.

Krupskii, P. and Joe, H. (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis*, 120:85–101.

Krupskii, P. and Joe, H. (2015). Structured factor copula models: Theory, inference and computation. *Journal of Multivariate Analysis*, 138:53 – 73. High-Dimensional Dependence and Copulas.

Krupskii, P. and Joe, H. (2022). Approximate likelihood with proxy variables for parameter estimation in high-dimensional factor copula models. *Statistical Papers*, 63:543–569.

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

Kurowicka, D. and Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapore.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15(3):209–225.

Lee, G. and Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12(3):237–255.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, 57(1):89–105.

Ma, Y. and Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):475–495.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):22.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71(4):713–732.

Maydeu-Olivares, A. and Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49(4):305–328.

McDonald, R. P. (1997). Normal ogive multidimensional model. In van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of modern item response theory*, New York. Springer.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *Annals of Statistics*, 37:3059–3097.

Montanari, A. and Viroli, C. (2010). A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, 37(3):473–487.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49:313–334.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3):391–411.

Moustaki, I. and Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, 101(474):644–653.

Mulaik, S. A. and Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal*, 4(3):193–211.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4):551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Nagler, T., Bumann, C., and Czado, C. (2019). Model selection in sparse high-dimensional vine copula models with an application to portfolio risk. *Journal of Multivariate Analysis*, 172:180–192. Dependence Models.

Nash, J. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Hilger, New York. 2nd edition.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, 2nd edition.

Nikoloulopoulos, A. K. (2013a). Copula-based models for multivariate discrete response data. In Durante, F., Härdle, W., and Jaworski, P., editors, *Copulae in Mathematical and Quantitative Finance*, volume vol 213, pages 231–249, Berlin, Heidelberg. Springer.

Nikoloulopoulos, A. K. (2013b). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, 143(11):1923–1937.

Nikoloulopoulos, A. K. (2016). Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic Environmental Research and Risk Assessment*, 30(2):493–505.

Nikoloulopoulos, A. K. (2017). A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Statistical Methods in Medical Research*, 26(5):2270–2286.

Nikoloulopoulos, A. K. and Joe, H. (2015). Factor copula models for item response data. *Psychometrika*, 80(1):126–150.

Nikoloulopoulos, A. K., Joe, H., and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics & Data Analysis*, 56:3659–3673.

Nikoloulopoulos, A. K. and Karlis, D. (2008). Copula model evaluation based on parametric bootstrap. *Computational Statistics & Data Analysis*, 52:3342–3353.

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.

North, D. C. and Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *The Journal of Economic History*, 49(4):803–832.

Olsson, F. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44:443–460.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

Panagiotelis, A., Czado, C., Joe, H., and Stöber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis*, 106:138–152.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401.

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, 12(4):338–353.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3(3):215–232.

Rabe-Hesketh, S. and Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, 57(4):1256–1263.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47(3):361–372.

Rijmen, F., Jeon, M., von Davier, M., and Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, 39(4):235–256.

Rizopoulos, D. and Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61(2):415–438.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17.

Schmid, J. and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61.

Schroeders, U., Kubera, F., and Gnambs, T. (2021). The structure of the Toronto Alexithymia Scale (TAS-20): A meta-analytic confirmatory factor analysis. *Assessment*, pages 1–18.

Shen, C. and Weissfeld, L. (2006). A copula model for repeated measurements with non-ignorable non-monotone missing outcome. *Statistics in medicine*, 25(14):2427–40.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–1399.

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3):237–247.

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

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC Press.

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

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

Stöber, J., Hong, H. G., Czado, C., and Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis*, 88:28–39.

Stroud, A. and Secrest, D. (1966). *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, NJ.

Takane, Y. and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408.

Tuliao, A. P., Klanecky, A. K., Landoy, B. V. N., and McChargue, D. E. (2020). Toronto Alexithymia Scale–20: examining 18 competing factor structure solutions in a U.S. sample and a Philippines sample. *Assessment*, 27(7):1515–1531.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge University Press.

Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3):185–201.

Wainer, H. and Thissen, D. (1996). How is reliability related to the quality of test scores? what is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1):22–29.

Wang, W.-C. and Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29(2):126–149.

Wedel, M. and Kamakura, W. A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530.

Williams, D. and Mulder, J. (2020). *BGGM: Bayesian Gaussian Graphical Models*. R package version 1.0.0.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3):187–213.

Yung, Y.-F., Thissen, D., and McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64(2):113–128.

Zenisky, A. L., Hambleton, R. K., and Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39(4):291–309.

Zilko, A. A. and Kurowicka, D. (2016). Copula in a multivariate mixed discrete–continuous model. *Computational Statistics & Data Analysis*, 103:28–55.

Appendix A

Package ‘FactorCopula’

Package 'FactorCopula'

June 4, 2022

Version 0.8

Date 2021-12-18

Title Factor, Bi-Factor and Second-Order Copula Models

Author Sayed H. Kadhem [aut, cre],
Aristidis K. Nikoloulopoulos [aut]

Maintainer Sayed H. Kadhem <s.kadhem@uea.ac.uk>

Depends R (>= 3.5.0), statmod, abind, utils, polycor, VineCopula, matlab

Description Estimation, model selection and goodness-of-fit of (1) factor copula models for mixed continuous and discrete data in Kadhem and Nikoloulopoulos (2021) <doi:10.1111/bmsp.12231>; (2) bi-factor and second-order copula models for item response data in Kadhem and Nikoloulopoulos (2021) <arXiv:2102.10660>.

License GPL (>= 2)

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-12-19 01:52:15 UTC

R topics documented:

`FactorCopula-package` *Factor, bi-factor and second-order copula models*

Description

Estimation, model selection and goodness-of-fit of (1) factor copula models for mixed continuous and discrete data in Kadhem and Nikoloulopoulos (2021a); (2) bi-factor and second-order copula models for item response data in Kadhem and Nikoloulopoulos (2021b).

Details

This package contains R functions for:

- diagnostics based on semi-correlations (Kadhem and Nikoloulopoulos, 2021a,b; Joe, 2014) to detect tail dependence or tail asymmetry;
- diagnostics to show that a dataset has a factor structure based on linear factor analysis (Kadhem and Nikoloulopoulos, 2021a,b ; Joe, 2014);
- estimation of the factor copula models in Krupskii and Joe (2013), Nikoloulopoulos and Joe (2015), and Kadhem and Nikoloulopoulos (2021a, 2021b);
- model selection of the factor copula models in Krupskii and Joe (2013), Nikoloulopoulos and Joe (2015) and Kadhem and Nikoloulopoulos (2021a, 2021b) using the heuristic algorithms in Kadhem and Nikoloulopoulos (2021a, 2021b) that automatically selects the bivariate parametric copula families that link the observed to the latent variables;
- goodness-of-fit of the factor copula models in Krupskii and Joe (2013), Nikoloulopoulos and Joe (2015) and Kadhem and Nikoloulopoulos (2021a, 2021b) using the M_2 statistic (Maydeu-Olivares and Joe, 2006). Note that the continuous and count data have to be transformed to ordinal.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman & Hall, London.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, **71**, 713–732. doi: 10.1007/s1133600512959.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021a) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021b) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

Krupskii, P. and Joe, H. (2013) Factor copula models for multivariate data. *Journal of Multivariate Analysis*, **120**, 85–101. doi: 10.1016/j.jmva.2013.05.001.

Nikoloulopoulos, A.K. and Joe, H. (2015) Factor copula models with item response data. *Psychometrika*, **80**, 126–150. doi: 10.1007/s1133601393874.

`discrepancy` *Diagnostics to detect a factor dependence structure*

Description

The diagnostic method in Joe (2014, pages 245-246) to show that each dataset has a factor structure based on linear factor analysis. The correlation matrix $\mathsf{R}_{\text{observed}}$ has been obtained based on the sample correlations from the bivariate pairs of the observed variables. These are the linear (when both variables are continuous), polychoric (when both variables are ordinal), and polyserial (when one variable is continuous and the other is ordinal) sample correlations among the observed variables. The resulting $\mathsf{R}_{\text{observed}}$ is generally positive definite if the sample size is not small enough; if not one has to convert it to positive definite. We calculate various measures of discrepancy between $\mathsf{R}_{\text{observed}}$ and $\mathsf{R}_{\text{model}}$ (the resulting correlation matrix of linear factor analysis), such as the maximum absolute correlation difference $D_1 = \max|\mathsf{R}_{\text{model}} - \mathsf{R}_{\text{observed}}|$, the average absolute correlation difference $D_2 = \text{avg}|\mathsf{R}_{\text{model}} - \mathsf{R}_{\text{observed}}|$, and the correlation matrix discrepancy measure $D_3 = \log\big(\det(\mathsf{R}_{\text{model}})\big) - \log\big(\det(\mathsf{R}_{\text{observed}})\big) + \text{tr}(\mathsf{R}_{\text{model}}^{-1}\mathsf{R}_{\text{observed}}) - d$.

Usage

```
discrepancy(cormat, n, f3)
```

Arguments

`cormat`	$\mathsf{R}_{\text{observed}}$.
`n`	Sample size.
`f3`	If TRUE, then the linear 3-factor analysis is fitted.

Value

A matrix with the calculated discrepancy measures for different number of factors.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman & Hall, London.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Examples

```
#-----------------------------------------------
#                    PE Data
#------------------            -----------------
data(PE)
#correlation
continuous.PE1 <- -PE[,1]
continuous.PE <- cbind(continuous.PE1, PE[,2])
```

```
u.PE <- apply(continuous.PE, 2, rank)/(nrow(PE)+1)
z.PE <- qnorm(u.PE)
categorical.PE <- data.frame(apply(PE[, 3:5], 2, factor))
nPE <- cbind(z.PE, categorical.PE)

#------------------------------------------------
# Discrepancy measures--------------------------
#------------------------------------------------
#correlation matrix for mixed data
cormat.PE <- as.matrix(polycor::hetcor(nPE, std.err=FALSE))
#discrepancy measures
out.PE = discrepancy(cormat.PE, n = nrow(nPE), f3 = FALSE)

#------------------------------------------------
#------------------------------------------------
#                    GSS Data
#-----------------            -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME,AGE)
continuous.GSS <- apply(continuous.GSS, 2, rank)/(nrow(GSS)+1)
z.GSS <- qnorm(continuous.GSS)
ordinal.GSS <- cbind(DEGREE,PINCOME,PDEGREE)
count.GSS <- cbind(CHILDREN,PCHILDREN)

# Transforming the count variables to ordinal
# count1 : CHILDREN
count1  = count.GSS[,1]
count1[count1 > 3] = 3

# count2: PCHILDREN
count2  = count.GSS[,2]
count2[count2 > 7] = 7

# Combining both transformed count variables
ncount.GSS = cbind(count1, count2)

# Combining ordinal and transformed count variables
categorical.GSS <- cbind(ordinal.GSS, ncount.GSS)
categorical.GSS <- data.frame(apply(categorical.GSS, 2, factor))

# combining continuous and categorical variables
nGSS = cbind(z.GSS, categorical.GSS)

#------------------------------------------------
# Discrepancy measures--------------------------
#------------------------------------------------
#correlation matrix for mixed data
cormat.GSS <- as.matrix(polycor::hetcor(nGSS, std.err=FALSE))
#discrepancy measures
out.GSS = discrepancy(cormat.GSS, n = nrow(nGSS), f3 = TRUE)
```

Description

Hoff (2007) analysed seven demographic variables of 464 male respondents to the 1994 General Social Survey. Of these seven, two were continuous (income and age of the respondents), three were ordinal with 5 categories (highest degree of the survey respondent, income and highest degree of respondent's parents), and two were count variables (number of children of the survey respondent and respondent's parents).

Usage

```
data(GSS)
```

Format

A data frame with 464 observations on the following 7 variables:

`INCOME` Income of the respondent in 1000s of dollars, binned into 21 ordered categories.

`DEGREE` Highest degree ever obtained (0:None, 1:HS, 2:Associates, 3:Bachelors, 4:Graduate).

`CHILDREN` Number of children of the survey respondent.

`PINCOME` Financial status of respondent's parents when respondent was 16 (on a 5-point scale).

`PDEGREE` Highest degree of the survey respondent's parents (0:None, 1:HS, 2:Associates, 3:Bachelors, 4:Graduate).

`PCHILDREN` Number of children of the survey respondent's parents - 1.

`AGE` Age of the respondents in years.

Source

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, **1**, 265–283.

M2.Factor *Goodness-of-fit of factor copula models for mixed data*

Description

The limited information M_2 statistic (Maydeu-Olivares and Joe, 2006) of factor copula models for mixed continuous and discrete data.

Usage

```
M2.1F(tcontinuous, ordinal, tcount, cpar, copF1, gl)
M2.2F(tcontinuous, ordinal, tcount, cpar, copF1, copF2, gl, SpC)
```

Arguments

`tcontinuous`	$n \times d_1$ matrix with the transformed continuous to ordinal reponse data, where n and d_1 is the number of observations and transformed continous variables, respectively.
`ordinal`	$n \times d_2$ matrix with the ordinal reponse data, where n and d_2 is the number of observations and ordinal variables, respectively.

tcount	$n \times d_3$ matrix with the transformed count to ordinal reponse data, where n and d_3 is the number of observations and transformed count variables, respectively.
cpar	A list of estimated copula parameters.
copF1	$(d_1 + d_2 + d_3)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 1st factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7,"BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
copF2	$(d_1 + d_2 + d_3)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 2nd factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7,"BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
gl	Gauss legendre quardrature nodes and weights.
SpC	Special case for the 2-factor copula model with BVN copulas. Select a bivariate copula at the 2nd factor to be fixed to independence. e.g. "SpC = 1" to set the first copula at the 2nd factor to independence.

Details

The M_2 statistic has been developed for goodness-of-fit testing in multidimensional contingency tables by Maydeu-Olivares and Joe (2006). Nikoloulopoulos and Joe (2015) have used the M_2 statistic to assess the goodness-of-fit of factor copula models for ordinal data. We build on the aforementioned papers and propose a methodology to assess the overall goodness-of-fit of factor copula models for mixed continuous and discrete responses. Since the M_2 statistic has been developed for multivariate ordinal data, we propose to first transform the continuous and count variables to ordinal and then calculate the M_2 statistic at the maximum likelihood estimate before transformation.

Value

A list containing the following components:

M2	The M_2 statistic which has a null asymptotic distribution that is χ^2 with $s - q$ degrees of freedom, where s is the number of univariate and bivariate margins that do not include the category 0 and q is the number of model parameters.
df	$s - q$.
p-value	The resultant p-value.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, **71**, 713–732. doi: 10.1007/s1133600512959.

Nikoloulopoulos, A.K. and Joe, H. (2015) Factor copula models with item response data. *Psychometrika*, **80**, 126–150. doi: 10.1007/s1133601393874.

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                     PE Data
#-----------------             -----------------
data(PE)
continuous.PE1 = -PE[,1]
continuous.PE2 = PE[,2]
continuous.PE <- cbind(continuous.PE1, continuous.PE2)

categorical.PE <- PE[, 3:5]
#-------------------------------------------------
#                   Estimation
#-----------------             -----------------
#----------------- One-factor  -----------------
# one-factor copula model
cop1f.PE <- c("joe", "joe", "rjoe", "joe", "gum")
est1factor.PE <- mle1factor(continuous.PE, categorical.PE,
                            count=NULL, copF1=cop1f.PE, gl, hessian = T)
#-------------------------------------------------
#                      M2
#-------------------------------------------------
#Transforming the continuous to ordinal data:
ncontinuous.PE = continuous2ordinal(continuous.PE, 5)
# M2 statistic for the one-factor copula model:

m2.1f.PE <- M2.1F(ncontinuous.PE, categorical.PE, tcount=NULL,
       cpar=est1factor.PE$cpar, copF1=cop1f.PE, gl)

#-------------------------------------------------
#                    GSS Data
#-----------------             -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME,AGE)
ordinal.GSS <- cbind(DEGREE,PINCOME,PDEGREE)
count.GSS <- cbind(CHILDREN,PCHILDREN)

#-------------------------------------------------
#                   Estimation
#-----------------             -----------------
# one-factor copula model
```

```
cop1f.GSS <- c("joe","2rjoe","bvt3","bvt3",
          "rgum","2rjoe","2rgum")
est1factor.GSS <- mle1factor(continuous.GSS, ordinal.GSS,
                       count.GSS, copF1=cop1f.GSS, gl, hessian = T)

#two-factor copula model
cop1.2f <- c("rgum","rjoe","bvn","1rjoe",
           "1rjoe","rjoe","gum")
cop2.2f <- c("gum","2rjoe","rjoe","gum",
           "bvt5","bvn","2rgum")
est2factor.GSS <- mle2factor(continuous.GSS, ordinal.GSS,
        count.GSS, copF1=cop1.2f, copF2=cop2.2f, gl, hessian = T)

#-------------------------------------------------
#                 Transformation
#-----------------            -----------------
# Transforming the continuous to ordinal data:

# continuous1: Income
continuous1 = as.integer(cut(continuous.GSS[,1],
c(0,10,19,29,40,100), include.lowest = T))
continuous1 = continuous1 - 1

# continuous2: AGE
continuous2 = as.integer(cut(continuous.GSS[,2] ,
c(0, 24, 44, 64, 100), include.lowest = T))
continuous2 = continuous2 - 1

# Combining the transformed continuous variables.
ncontinuous.GSS <- cbind(continuous1, continuous2)

#---------------- COUNT VARIABLE -----------------
# count1 : CHILDREN
count1  = count.GSS[,1]
count1[count1 > 3] = 3

# count2: PCHILDREN
count2  = count.GSS[,2]
count2[count2 > 7] = 7

# Combining both transformed count variables
ncount.GSS = cbind(count1, count2)

#-------------------------------------------------
#                       M2
#-------------------------------------------------
# M2 statistic for the one-factor copula model:

m2.1f.GSS <- M2.1F(ncontinuous.GSS, ordinal.GSS,
                     ncount.GSS, cpar = est1factor.GSS$cpar,
                     copF1 = cop1f.GSS, gl)

#-------------------------------------------------
# M2 statistic for the two-factor copula model:

m2.2f.GSS <- M2.2F(ncontinuous.GSS, ordinal.GSS, ncount.GSS,
                     cpar = est2factor.GSS$cpar, copF1 = cop1.2f,
```

```
copF2 = cop2.2f, gl)
```

`M2.StructuredFactor` *Goodness-of-fit of bi-factor and second-order copula models for item response data*

Description

The limited information M_2 statistic (Maydeu-Olivares and Joe, 2006) of bi-factor and second-order copula models for item response data.

Usage

```
M2Bifactor(y,cpar, copnames1, copnames2, gl, ngrp, grpsize)
M2Second_order(y,cpar, copnames1, copnames2, gl, ngrp, grpsize)
```

Arguments

`y`	$n \times d$ matrix with the ordinal reponse data, where n and d is the number of observations and variables, respectively.
`cpar`	A list of estimated copula parameters.
`copnames1`	**For the bi-factor copula**: d-vector with the names of bivariate copulas that link each of the oberved variabels with the common factor. **For the second-order factor copula**: G-vector with the names of bivariate copulas that link the each of the group-specific factors with the common factor, where G is the number of groups of items. Choices are "bvn" for BVN, "bvtν" with $\nu = \{2, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.
`copnames2`	**For the bi-factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. **For the second-order factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{2, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.
`gl`	Gauss legendre quardrature nodes and weights.
`ngrp`	number of non-overlapping groups.
`grpsize`	vector indicating the size for each group, e.g., c(4,4,4) indicating four items in all three groups.

Details

The M_2 statistic has been developed for goodness-of-fit testing in multidimensional contingency tables by Maydeu-Olivares and Joe (2006). We use the M_2 to assess the overall fit for the bi-factor and second-order copula models for item resposne data (Kadhem & Nikoloulopoulos, 2021).

Value

A list containing the following components:

M2	The M_2 statistic which has a null asymptotic distribution that is χ^2 with $s - q$ degrees of freedom, where s is the number of univariate and bivariate margins that do not include the category 0 and q is the number of model parameters.
df	$s - q$.
p-value	The resultant p-value.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, **71**, 713–732. doi: 10.1007/s1133600512959.

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                    TAS Data
#-----------------             -----------------
data(TAS)
grp1=c(1,3,6,7,9,13,14)
grp2=c(2,4,11,12,17)
grp3=c(5,8,10,15,16,18,19,20)
#Rearrange items within testlets
ydat=TAS[,c(grp1,grp2,grp3)]

d=ncol(ydat);d
n=nrow(ydat);n

#size of each group
g1=length(grp1)
g2=length(grp2)
g3=length(grp3)

grpsize=c(g1,g2,g3)#group size
#number of groups
ngrp=length(grpsize)

#-------------------------------------------------
#                      M2
#-------------------------------------------------
#BI-FACTOR
```

```
tauX0 = c(0.49,0.16,0.29,0.09,0.47,0.49,0.30,
          0.46,0.41,0.33,0.29,0.24,0.10,0.16,
          0.14,0.12,0.03,0.03,0.10,0.10)
tauXg = c(0.09,0.37,0.23,0.53,0.24,0.32,0.27,
          0.53,0.58,0.20,0.23,0.25,0.34,0.33,
          0.30,0.19,0.24,0.29,0.43,0.26)
copX0 = rep("bvt2", d)
copXg = c(rep("rgum", g1), rep("bvt3", g2+g3))
#converting taus to cpars
cparX0=mapply(function(x,y) tau2par(x,y),x=copX0,y=tauX0)
cparXg=mapply(function(x,y) tau2par(x,y),x=copXg,y=tauXg)
cpar=c(cparX0,cparXg)

m2_Bifactor = M2Bifactor(y=ydat, cpar, copX0, copXg, gl, ngrp, grpsize)

#SECOND-ORDER
tauX0Xg=c(0.60,0.74,0.18)
tauXgY=c(0.48,0.23,0.34,0.25,0.51,0.56,0.37,0.64,0.57,
0.37,0.35,0.32,0.33,0.33,0.29,0.23,0.23,0.25,0.39,0.28)
cparX0Xg=tau2par("bvn",tauX0Xg)
cparXgY=tau2par("bvn",tauXgY)

cpar=c(cparX0Xg,cparXgY)
copX0Xg = rep("bvn", ngrp)
copXgY = rep("bvn", g1+g2+g3)
m2_Second_order =  M2Second_order(y=ydat,cpar, copX0Xg, copXgY, gl, ngrp, grpsize)
```

`mapping` *Mapping of Kendall's tau and copula parameter*

Description

Bivariate copulas: mapping of Kendall's tau and copula parameter.

Usage

```
par2tau(copulaname, cpar)
tau2par(copulaname, tau)
```

Arguments

`copulaname`	Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
`cpar`	Copula parameter(s).
`tau`	Kendall's tau.

Value

Kendall's tau or copula parameter.

References

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

Joe H (2014) *Dependence Modeling with Copulas*. Chapman & Hall, London.

Joe H (2014) *CopulaModel: Dependence Modeling with Copulas*. Software for book: *Dependence Modeling with Copulas*, Chapman & Hall, London 2014.

Examples

```
# 1-param copulas
#BVN copula
cpar.bvn = tau2par("bvn", 0.5)
tau.bvn = par2tau("bvn", cpar.bvn)

#Frank copula
cpar.frk = tau2par("frk", 0.5)
tau.frk = par2tau("frk", cpar.frk)

#Gumbel copula
cpar.gum = tau2par("gum", 0.5)
tau.gum = par2tau("gum", cpar.gum)

#Joe copula
cpar.joe = tau2par("joe", 0.5)
tau.joe = par2tau("joe", cpar.joe)

# 2-param copulas
#BB1 copula
tau.bb1 = par2tau("bb1", c(0.5,1.5))

#BB7 copula
tau.bb7 = par2tau("bb7", c(1.5,1))

#BB8 copula
tau.bb8 = par2tau("bb8", c(3,0.8))

#BB10 copula
tau.bb10 = par2tau("bb10", c(3,0.8))
```

mle.Factor *Maximum likelihood estimation of factor copula models for mixed data*

Description

We use a two-stage etimation approach toward the estimation of factor copula models for mixed continuous and discrete data.

Usage

```
mle1factor(continuous, ordinal, count, copF1, gl, hessian, print.level)
mle2factor(continuous, ordinal, count, copF1, copF2, gl, hessian, print.level)
mle2factor.bvn(continuous, ordinal, count, copF1, copF2, gl, SpC, print.level)
```

Arguments

continuous	$n \times d_1$ matrix with the continuous reponse data, where n and d_1 is the number of observations and continous variables, respectively.
ordinal	$n \times d_2$ matrix with the ordinal reponse data, where n and d_2 is the number of observations and ordinal variables, respectively.
count	$n \times d_3$ matrix with the count reponse data, where n and d_3 is the number of observations and count variables, respectively.
copF1	$(d_1 + d_2 + d_3)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 1st factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
copF2	$(d_1 + d_2 + d_3)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 2nd factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
gl	Gauss legendre quardrature nodes and weights.
SpC	Special case for the 2-factor copula model with BVN copulas. Select a bivariate copula at the 2nd factor to be fixed to independence. e.g. "SpC = 1" to set the first copula at the 2nd factor to independence.
hessian	If TRUE, the hessian of the negative log-likelihood is calculated during the minimization process.
print.level	Determines the level of printing which is done during the minimization process; same as in nlm.

Details

Estimation is achieved by maximizing the joint log-likelihood over the copula parameters with the univariate parameters/distributions fixed as estimated at the first step of the proposed two-step estimation approach.

Value

A list containing the following components:

cutpoints	The estimated univariate cutpoints (fitting the univariate probit model).
negbinest	The estimated univariate parametes for the count responses (fitting the negative binomial distribution).

loglik	The maximized joint log-likelihood.
cpar	Estimated copula parameters in a list form.
taus	The estimated copula parameters in Kendall's tau scale.
SEs	The SEs of the Kendall's tau estimates.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Krupskii, P. and Joe, H. (2013) Factor copula models for multivariate data. *Journal of Multivariate Analysis*, **120**, 85–101. doi: 10.1016/j.jmva.2013.05.001.

Nikoloulopoulos, A.K. and Joe, H. (2015) Factor copula models with item response data. *Psychometrika*, **80**, 126–150. doi: 10.1007/s11336-013-9393-4.

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                     PE Data
#-----------------             -----------------
data(PE)
continuous.PE1 = -PE[,1]
continuous.PE2 = PE[,2]
continuous.PE <- cbind(continuous.PE1, continuous.PE2)

categorical.PE <- PE[, 3:5]
#-------------------------------------------------
#                   Estimation
#-----------------             -----------------
#----------------- One-factor  -----------------
# one-factor copula model
cop1f.PE <- c("joe", "joe", "rjoe", "joe", "gum")
est1factor.PE <- mle1factor(continuous.PE, categorical.PE,
                            count=NULL, copF1=cop1f.PE, gl, hessian = T)
est1factor.PE
#-------------------------------------------------
#-------------------------------------------------
#                     GSS Data
#-----------------             -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME, AGE)
ordinal.GSS <- cbind(DEGREE, PINCOME, PDEGREE)
count.GSS <- cbind(CHILDREN, PCHILDREN)
```

```
#-------------------------------------------------
#                   Estimation
#-----------------              -----------------
#------------------ One-factor  -----------------
# one-factor copula model
cop1f.GSS <- c("joe","2rjoe","bvt3","bvt3",
          "rgum","2rjoe","2rgum")
est1factor.GSS <- mle1factor(continuous.GSS, ordinal.GSS,
                       count.GSS, copF1 = cop1f.GSS, gl, hessian = T)

#------------------ Two-factor  -----------------
# two-factor copula model
cop1.2f <- c("rgum","rjoe","bvn","1rjoe",
           "1rjoe","rjoe","gum")
cop2.2f <- c("gum","2rjoe","rjoe","gum",
           "bvt5","bvn","2rgum")
est2factor.GSS <- mle2factor(continuous.GSS, ordinal.GSS,
          count.GSS, copF1 = cop1.2f, copF2 = cop2.2f, gl, hessian = T)
```

`mle.StructuredFactor` *Maximum likelihood estimation of the bi-factor and second-order copula models for item response data*

Description

We approach the estimation of the bi-factor and second-order copula models for item response data with the IFM method of Joe (2005).

Usage

```
mleBifactor(y, copnames1, copnames2, gl, ngrp, grpsize,
hessian, print.level)
mleSecond_order(y, copnames1, copnames2, gl, ngrp, grpsize,
hessian, print.level)
```

Arguments

`y`	$n \times d$ matrix with the item reponse data, where n and d is the number of observations and variables, respectively.
`copnames1`	**For the bi-factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the common factor. **For the second-order factor copula**: G-vector with the names of bivariate copulas that link the each of the group-specific factors with the common factor, where G is the number of groups of items. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.

copnames2	**For the bi-factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. **For the second-order factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.
gl	Gauss legendre quardrature nodes and weights.
ngrp	number of non-overlapping groups.
grpsize	vector indicating the size for each group, e.g., c(4,4,4) indicating four items in all three groups.
hessian	If TRUE, the hessian of the negative log-likelihood is calculated during the minimization process.
print.level	Determines the level of printing which is done during the minimization process; same as in nlm.

Details

Estimation is achieved by maximizing the joint log-likelihood over the copula parameters with the univariate cutpoints fixed as estimated at the first step of the proposed two-step estimation approach.

Value

A list containing the following components:

cutpoints	The estimated univariate cutpoints (fitting the univariate probit model).
taus	The estimated copula parameters in Kendall's tau scale.
SEs	The SEs of the Kendall's tau estimates.
loglik	The maximized joint log-likelihood.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Joe, H. (2005) Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, **94**, 401–419. doi: 10.1016/j.jmva.2004.06.003.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

Examples

```
#-----------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-----------------------------------------------
#                    TAS Data
#-----------------             -----------------
```

```
data(TAS)
grp1=c(1,3,6,7,9,13,14)
grp2=c(2,4,11,12,17)
grp3=c(5,8,10,15,16,18,19,20)
#Rearrange items within testlets
ydat=TAS[,c(grp1,grp2,grp3)]

d=ncol(ydat);d
n=nrow(ydat);n

#size of each group
g1=length(grp1)
g2=length(grp2)
g3=length(grp3)

grpsize=c(g1,g2,g3)#group size
#number of groups
ngrp=length(grpsize)

#BI-FACTOR
copX0 = rep("bvt2", d)
copXg = c(rep("rgum", g1), rep("bvt3", g2+g3))
mle_Bifactor =  mleBifactor(y = ydat, copX0, copXg, gl, ngrp, grpsize, hessian=F, print.level=2)

#SECOND-ORDER
copX0Xg = rep("bvt5", ngrp)
copXgY = c(rep("bvt3", g1), rep("bvt2", g2+g3))
mle_Second_order = mleSecond_order(y = ydat, copX0Xg,
                                   copXgY, gl, ngrp, grpsize,
                                   hessian=F, print.level=2)
```

PE *Political-economic risk of 62 countries for the year 1987*

Description

Quinn (2004) used 5 mixed variables, namely the continuous variable black-market premium in each country (used as a proxy for illegal economic activity), the continuous variable productivity as measured by real gross domestic product per worker in 1985 international prices, the binary variable independence of the national judiciary (1 if the judiciary is judged to be independent and 0 otherwise), and the ordinal variables measuring the lack of expropriation risk and lack of corruption.

Usage

```
data(PE)
```

Format

A data frame with 62 observations (countries) on the following 5 variables:

`BM` Black-market premium.

`GDP` Gross domestic product.

`IJ` Independent judiciary.

`XPR` Lack of expropriation risk.

`CPR` Lack of corruption.

Source

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, **12**, 338–353.

`rFactor` *Simulation of factor copula models for mixed continuous and discrete data*

Description

Simulating dependent standard uniform and ordinal response data from factor copula models.

Usage

```
r1factor(n, d1, d2, categ, theta, copF1)
r2factor(n, d1, d2, categ, theta, delta, copF1, copF2)
```

Arguments

`n`	Sample size.
`d1`	Number of standard uniform variables.
`d2`	Number of ordinal variables.
`categ`	A vector of categories for the ordinal variables.
`theta`	Copula parameters for the 1st factor.
`delta`	Copula parameters for the 2nd factor.
`copF1`	$(d_1 + d_2)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 1st factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.
`copF2`	$(d_1 + d_2)$-vector with the names of bivariate copulas that link the each of the oberved variabels with the 2nd factor. Choices are "bvn" for BVN, "bvt[ν]" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.

Value

Data matrix of dimension $n \times d$, where n is the sample size, and $d = d_1 + d_2$ is the total number of variables.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Examples

```
# ------------------------------------------------
# ------------------------------------------------
#              One-factor copula model
# ------------------------------------------------
# ------------------------------------------------
#Sample size ----------------------------------------
n = 100

#Continuous Variables  ------------------------------
d1 = 5

#Ordinal Variables  ---------------------------------
d2 = 3

#Categories for ordinal  ----------------------------
categ = c(3,4,5)

#Copula parameters  ---------------------------------
theta = rep(2, d1+d2)

#Copula names  --------------------------------------
copnamesF1 = rep("gum", d1+d2)

#----------------- Simulating data ------------------
datF1 = r1factor(n, d1=d1, d2=d2, categ, theta, copnamesF1)

#------------  Plotting continuous data -------------
pairs(qnorm(datF1[, 1:d1]))

# ------------------------------------------------
# ------------------------------------------------
#              Two-factor copula model
# ------------------------------------------------
# ------------------------------------------------
#Sample size ----------------------------------------
n = 100

#Continuous Variables  ------------------------------
d1 = 5

#Ordinal Variables  ---------------------------------
d2 = 3
```

```
#Categories for ordinal  ----------------------------
categ = c(3,4,5)

#Copula parameters  ---------------------------------
theta = rep(2.5, d1+d2)
delta = rep(1.5, d1+d2)

#Copula names  --------------------------------------
copnamesF1 = rep("gum", d1+d2)
copnamesF2 = rep("gum", d1+d2)

#----------------- Simulating data ------------------
datF2 = r2factor(n, d1=d1, d2=d2, categ, theta, delta,
                copnamesF1, copnamesF2)

#-----------------  Plotting  data ------------------
pairs(qnorm(datF2[,1:d1]))
```

`rStructuredFactor` *Simulation of bi-factor and second-order copula models for item response data*

Description

Simulating dependent item response data from the bi-factor and second-order copula models for item response data.

Usage

```
rBifactor(n, d, grpsize, categ, copnames1,copnames2, theta1, theta2)
rSecond_order(n, d, grpsize, categ, copnames1, copnames2, theta1, theta2)
```

Arguments

`n`	Sample size.
`d`	Number of observed variables/items.
`grpsize`	vector indicating the size for each group, e.g., c(4,4,4) indicating four items in all three groups.
`categ`	A vector of categories for the observed variables/items.
`theta1`	**For the bi-factor model**: copula parameter vector of size d for items with the common factor. **For the second-order copulas**: copula parameter vector of size G for the common factor and group-specific factors.
`theta2`	**For the bi-factor model**: copula parameter vector of size d for items with the group-specific factor. **For the second-order copulas**: copula parameter vector of size d for items with the group-specific factor.
`copnames1`	**For the bi-factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the common factor. **For the second-order factor copula**: G-vector with the names of bivariate copulas that link the each of the group-specific factors with the common factor, where G is the number

of groups of items. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.

`copnames2` **For the bi-factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. **For the second-order factor copula**: d-vector with the names of bivariate copulas that link the each of the oberved variabels with the group-specific factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.

Value

Data matrix of dimension $n \times d$, where n is the sample size, and d is the total number of observed variables/items.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

Examples

```
# ----------------------------------------------------
# ----------------------------------------------------
#Sample size
n = 500

#Ordinal Variables  ---------------------------------
d = 9
grpsize=c(3,3,3)
ngrp=length(grpsize)

#Categories for ordinal  ----------------------------
categ = rep(3,d)

# ----------------------------------------------------
# ----------------------------------------------------
#              Bi-factor copula model
# ----------------------------------------------------
# ----------------------------------------------------
#Copula parameters
theta = rep(2.5, d)
delta = rep(1.5, d)

#Copula names
copulanames1 = rep("gum", d)
copulanames2 = rep("gum", d)
```

```
#----------------- Simulating data ------------------
data_Bifactor = rBifactor(n, d, grpsize, categ, copulanames1,
copulanames2, theta, delta)

# ---------------------------------------------------
# ---------------------------------------------------
#              Second-order copula model
# ---------------------------------------------------
# ---------------------------------------------------
#Copula parameters
theta= rep(1.5, ngrp)
delta = rep(2.5, d)

#Copula names
copulanames1 = rep("gum", ngrp)
copulanames2 = rep("gum", d)

data_Second_order = rSecond_order(n, d, grpsize, categ,
copulanames1, copulanames2, theta, delta)
```

`Select.Factor` *Model selection of the factor copula models for mixed data*

Description

A heuristic algorithm that automatically selects the bivariate parametric copula families that link the observed to the latent variables.

Usage

```
select1F(continuous, ordinal, count, copnamesF1, gl)
select2F(continuous, ordinal, count, copnamesF1, copnamesF2, gl)
```

Arguments

`continuous`	$n \times d_1$ matrix with the continuous reponse data, where n and d_1 is the number of observations and continous variables, respectively.
`ordinal`	$n \times d_2$ matrix with the ordinal reponse data, where n and d_2 is the number of observations and ordinal variables, respectively.
`count`	$n \times d_3$ matrix with the count reponse data, where n and d_3 is the number of observations and count variables, respectively.
`copnamesF1`	A vector with the names of possible candidates of bivariate copulas that link the each of the oberved variabels with the 1st factor. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.

`copnamesF2` A list with the names of possible candidates of bivariate copulas that link the each of the oberved variabels with the 1st and 2nd factors. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel, "joe" for Joe, "rjoe" for reflected Joe, "1rjoe" for 1-reflected Joe, "2rjoe" for 2-reflected Joe, "BB1" for BB1, "rBB1" for reflected BB1, "BB7" for BB7, "rBB7" for reflected BB7, "BB8" for BB8, "rBB8" for reflected BB8, "BB10" for BB10, "rBB10" for reflected BB10.

`gl` Gauss legendre quardrature nodes and weights.

Details

The linking copulas at each factor are selected with a sequential algorithm under the initial assumption that linking copulas are Frank, and then sequentially copulas with non-tail quadrant independence are assigned to any of pairs where necessary to account for tail asymmetry (discrete data) or tail dependence (continuous data).

Value

A list containing the following components:

`''1st factor''` The selected bivariate linking copulas for the 1st factor.

`''2nd factor''` The selected bivariate linking copulas for the 2nd factor.

`AIC` Akaike information criterion.

`taus` The estimated copula parameters in Kendall's tau scale.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Examples

```
#------------------------------------------------
#                   Estimation
#-----------------              -----------------
# Setting quadreture points
nq<-25
gl<-gauss.quad.prob(nq)
#------------------------------------------------
#                     PE Data
#-----------------              -----------------
data(PE)
continuous.PE1 = -PE[,1]
continuous.PE <- cbind(continuous.PE1, PE[,2])
categorical.PE <- PE[, 3:5]
```

```
#------------------ One-factor  -----------------
# listing the possible copula candidates:
d <- ncol(PE)
copulasF1 <- rep(list(c("bvn", "bvt3", "bvt5", "frk", "gum",
"rgum", "rjoe","joe", "1rjoe","2rjoe", "1rgum","2rgum")), d)
out1F.PE <- select1F(continuous.PE, categorical.PE,
count=NULL, copulasF1, gl)

#-------------------------------------------------
#-------------------------------------------------
#                     GSS Data
#-----------------             -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME, AGE)
ordinal.GSS <- cbind(DEGREE, PINCOME, PDEGREE)
count.GSS <- cbind(CHILDREN, PCHILDREN)

#------------------ One-factor  -----------------
# listing the possible copula candidates:
d <- ncol(GSS)
copulasF1 <- rep(list(c("bvn", "bvt3", "bvt5", "frk", "gum",
"rgum", "rjoe","joe", "1rjoe","2rjoe", "1rgum","2rgum")), d)
out1F.GSS <- select1F(continuous.GSS, ordinal.GSS, count.GSS, copulasF1, gl)

#------------------ two-factor  -----------------
# listing the possible copula candidates:
copulasF1 = copulasF2 = rep(list(c("bvn", "bvt3", "bvt5", "frk",
"gum", "rgum", "rjoe","joe", "1rjoe","2rjoe", "1rgum","2rgum")), d)
out2F.GSS <- select2F(continuous.GSS, ordinal.GSS,
count.GSS, copulasF1, copulasF2, gl)
```

`Select.StructuredFactor` *Model selection of the bi-factor and second-order copula models for item response data*

Description

A heuristic algorithm that automatically selects the bivariate parametric copula families for the bi-factor and second-order copula models for item response data.

Usage

```
selectBifactor(y, grpsize, copnames, gl)
selectSecond_order(y, grpsize, copnames, gl)
```

Arguments

`y`	$n \times d$ matrix with the item reponse data, where n and d is the number of observations and variables, respectively.

`grpsize` vector indicating the size for each group, e.g., c(4,4,4) indicating four items in all three groups.

`copnames` A vector with the names of possible candidates of bivariate copulas to be selected for the bi-factor and second-order copula models for item response data. Choices are "bvn" for BVN, "bvtν" with $\nu = \{1, \ldots, 9\}$ degrees of freedom for t-copula, "frk" for Frank, "gum" for Gumbel, "rgum" for reflected Gumbel, "1rgum" for 1-reflected Gumbel, "2rgum" for 2-reflected Gumbel.

`gl` Gauss legendre quardrature nodes and weights.

Details

The linking copulas at each factor are selected with a sequential algorithm under the initial assumption that linking copulas are BVN, and then sequentially copulas with tail dependence are assigned to any of pairs where necessary to account for tail asymmetry.

Value

A list containing the following components:

`''common factor''` The selected bivariate linking copulas for the common factor (Bi-factor: copulas link items with the common factor. Second-order: copulas link group-specific factors with the common factor).

`''group-specific factor''` The selected bivariate linking copulas for the items with group-speicifc factors.

`log-likelihood` The maximized joint log-likelihood.

`taus` The estimated copula parameters in Kendall's tau scale.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                     TAS Data
#------------------             -----------------
data(TAS)
grp1=c(1,3,6,7,9,13,14)
grp2=c(2,4,11,12,17)
grp3=c(5,8,10,15,16,18,19,20)
ydat=TAS[,c(grp1,grp2,grp3)]
```

```
#size of each group
g1=length(grp1)
g2=length(grp2)
g3=length(grp3)
grpsize=c(g1,g2,g3)

# listing the possible copula candidates:
copnames=c("bvn", "bvt2", "bvt3",
"gum", "rgum")

Bifactor_model = selectBifactor(ydat, grpsize, copnames, gl)
Second_order_model = selectSecond_order(ydat, grpsize, copnames, gl)
```

`semicorr` *Diagnostics to detect tail dependence or tail asymmetry.*

Description

The sample versions of the correlation ρ_N, upper semi-correlation ρ_N^+ (correlation in the joint upper quadrant) and lower semi-correlation ρ_N^- (correlation in the joint lower quadrant). These are sample linear (when both variables are continuous), polychoric (when both variables are ordinal), and polyserial (when one variable is continuous and the other is ordinal) correlations.

Usage

```
semicorr(dat,type)
```

Arguments

`dat`	Data frame of mixed continuous and ordinal data.
`type`	A vector with 1's for the location of continuous data and 2's for the location of ordinal data.

Value

A matrix containing the following components for semicorr():

`rho`	ρ_N.
`lrho`	ρ_N^-.
`urho`	ρ_N^+.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman and Hall/CRC.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

Examples

```
#-------------------------------------------------
#                     PE Data
#-----------------             -----------------
data(PE)
#correlation
continuous.PE1 <- -PE[,1]
continuous.PE <- cbind(continuous.PE1, PE[,2])
categorical.PE <- data.frame(apply(PE[, 3:5], 2, factor))
nPE <- cbind(continuous.PE, categorical.PE)

#-------------------------------------------------
# Semi-correlations-----------------------------
#-------------------------------------------------
# Exclude the dichotomous variable
sem.PE = nPE[,-3]
semicorr.PE = semicorr(dat = sem.PE, type = c(1,1,2,2))
#-------------------------------------------------
#-------------------------------------------------
#                    GSS Data
#-----------------             -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME,AGE)
ordinal.GSS <- cbind(DEGREE,PINCOME,PDEGREE)
count.GSS <- cbind(CHILDREN,PCHILDREN)

# Transforming the COUNT variables to ordinal
# count1 : CHILDREN
count1  = count.GSS[,1]
count1[count1 > 3] = 3

# count2: PCHILDREN
count2  = count.GSS[,2]
count2[count2 > 7] = 7

# Combining both transformed count variables
ncount.GSS = cbind(count1, count2)

# Combining ordinal and transformed count variables
categorical.GSS <- cbind(ordinal.GSS, ncount.GSS)
categorical.GSS <- data.frame(apply(categorical.GSS, 2, factor))

# combining continuous and categorical variables
nGSS = cbind(continuous.GSS, categorical.GSS)
#-------------------------------------------------
# Semi-correlations-----------------------------
#-------------------------------------------------
semicorr.GSS = semicorr(dat = nGSS, type = c(1, 1, rep(2,5)))
```

TAS *Toronto Alexithymia Scale (TAS)*

Description

The Toronto Alexithymia Scale is the most utilized measure of alexithymia in empirical research and is composed of $d = 20$ items that can be subdivided into $G = 3$ non-overlapping groups: $d_1 = 7$ items to assess difficulty identifying feelings (DIF), $d_2 = 5$ items to assess difficulty describing feelings (DDF) and $d_3 = 8$ items to assess externally oriented thinking (EOT). Students were 17 to 25 years old and 58% of them were female and 42% were male. They were asked to respond to each item using one of K = 5 categories: "1 = completely disagree", "2 = disagree", "3 = neutral", "4 = agree", "5 = completely agree".

Usage

```
data(TAS)
```

Format

A data frame with 1925 observations on the following 20 items:

`DIF` items: 1,3,6,7,9,13,14.

`DDF` items: 2,4,11,12,17.

`EOT` items: 5,8,10,15,16,18,19,20.

Source

Briganti, G. and Linkowski, P. (2020). Network approach to items and domains from the toronto alexithymia scale. *Psychological Reports*, **123**, 2038–2052.

Williams, D. and Mulder, J. (2020). *BGGM: Bayesian Gaussian Graphical Models*. R package version 1.0.0.

transformation *Continuous/count to ordinal responses*

Description

Transforming a continuous/count to ordinal variable with K categories.

Usage

```
continuous2ordinal(continuous, categ)
count2ordinal(count, categ)
```

Arguments

`continuous`	Matrix of continuous data.
`count`	Matrix of count data.
`categ`	The number of categories K.

Examples

```
#-----------------------------------------------
#                     PE Data
#------------------             -----------------
data(PE)
continuous.PE <- PE[, 1:2]

#Transforming the continuous to ordinal data :
tcontinuous = continuous2ordinal(continuous.PE, categ=5)
table(tcontinuous)

#Transforming the count to ordinal data:
set.seed(12345)
count.data = rpois(1000, 3)
tcount = count2ordinal(count.data, 5)
table(tcount)
```

Vuong.Factor *Vuong's test for the comparison of factor copula models*

Description

Vuong (1989)'s test for the comparison of non-nested factor copula models for mixed data. We compute the Vuong's test between the factor copula model with BVN copulas (that is the standard factor model) and a competing factor copula model to reveal if the latter provides better fit than the standard factor model.

Usage

```
vuong.1f(cpar.bvn, cpar, copF1, continuous, ordinal, count, gl, param)
vuong.2f(cpar.bvn, cpar, copF1, copF2, continuous, ordinal, count, gl, param)
```

Arguments

cpar.bvn	copula parameters of the factor copula model with BVN copulas.
cpar	copula parameters of the competing factor copula model.
copF1	copula names for the first factor of the competing factor copula model.
copF2	copula names for the second factor of the competing factor copula model.
continuous	matrix of continuous data.
ordinal	matrix of ordinal data.
count	matrix of count data.
gl	gauss-legendre quardature points.
param	parameterization of estimated copula parameters. If FALSE, then cpar are the actual copula parameters without any transformation/reparamterization.

Value

A vector containing the following components:

z	the test statistic.
p.value	the p-value.
CI.left	lower/left endpoint of 95% confidence interval.
CI.right	upper/right endpoint of 95% confidence interval.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*, **74**, 365–403. doi: 10.1111/bmsp.12231.

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

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                     PE Data
#-----------------            -----------------
data(PE)
continuous.PE1 = -PE[,1]
continuous.PE2 = PE[,2]
continuous.PE <- cbind(continuous.PE1, continuous.PE2)
categorical.PE <- PE[, 3:5]
d <- ncol(PE)
#-------------------------------------------------
#                   Estimation
#-----------------            -----------------
# factor copula model with BVN copulas
cop1f.PE.bvn <- rep("bvn", d)
PE.bvn1f <- mle1factor(continuous.PE, categorical.PE,
count=NULL, copF1=cop1f.PE.bvn, gl, hessian = T)

# Selected factor copula model
cop1f.PE <- c("joe", "joe", "rjoe", "joe", "gum")
PE.selected1f <- mle1factor(continuous.PE, categorical.PE,
count=NULL, copF1=cop1f.PE, gl, hessian = T)
#-------------------------------------------------
#                   Vuong's test
#-----------------            -----------------
v1f.PE.selected <- vuong.1f(PE.bvn1f$cpar$f1,
PE.selected1f$cpar$f1,cop1f.PE, continuous.PE,
categorical.PE, count=NULL, gl, param=F)
```

```
#-----------------------------------------------
#-----------------------------------------------
#                   GSS Data
#------------------            -----------------
data(GSS)
attach(GSS)
continuous.GSS <- cbind(INCOME,AGE)
ordinal.GSS <- cbind(DEGREE,PINCOME,PDEGREE)
count.GSS <- cbind(CHILDREN,PCHILDREN)
d <- ncol(GSS)

#-----------------------------------------------
#                  Estimation
#------------------            -----------------
# factor copula model with BVN copulas
# one-factor copula model
cop1f.GSS.bvn <- rep("bvn", d)
GSS.bvn1f <- mle1factor(continuous.GSS, ordinal.GSS,
count.GSS, copF1 = cop1f.GSS.bvn, gl, hessian = T)

# two-factor copula model
cop1f.GSS.bvn = cop2f.GSS.bvn = rep("bvn", d)
GSS.bvn2f <- mle2factor.bvn(continuous.GSS, ordinal.GSS,
count.GSS, copF1 = cop1f.GSS.bvn, copF2 = cop2f.GSS.bvn, gl, SpC =7)

# Selected factor copula  model
# one-factor copula model
cop1f.GSS <- c("joe","2rjoe","bvt3","bvt3",
               "rgum","2rjoe","2rgum")
GSS.selected1f <- mle1factor(continuous.GSS, ordinal.GSS,
count.GSS, copF1 = cop1f.GSS, gl, hessian = T)

# two-factor copula model
cop2f1.GSS <- c("rgum","rjoe","bvn","1rjoe","1rjoe","rjoe","gum")
cop2f2.GSS <- c("gum","2rjoe","rjoe","gum","bvt5","bvn","2rgum")
GSS.selected2f <- mle2factor(continuous.GSS, ordinal.GSS,
count.GSS, copF1 = cop2f1.GSS, copF2 = cop2f2.GSS, gl, hessian = T)

#-----------------------------------------------
#                  Vuong's test
#------------------            -----------------
#1-factor
v1f.GSS.selected <- vuong.1f(GSS.bvn1f$cpar$f1,
GSS.selected1f$cpar$f1, cop1f.GSS, continuous.GSS,
ordinal.GSS, count=count.GSS, gl, param=F)

#2-factor
v2f.GSS.selected <- vuong.2f(GSS.bvn2f$cpar,
GSS.selected2f$cpar, cop2f1.GSS, cop2f2.GSS,
continuous.GSS, ordinal.GSS, count=count.GSS, gl, param=F)
```

`Vuong.StructuredFactor` *Vuong's test for the comparison of bi-factor and second-order copula models*

Description

The Vuong's test (Vuong,1989) is the sample version of the difference in Kullback-Leibler divergence between two models and can be used to differentiate two parametric models which could be non-nested. For the Vuong's test we provide the 95% confidence interval of the Vuong's test statistic (Joe, 2014, page 258). If the interval does not contain 0, then the best fitted model according to the AICs is better if the interval is completely above 0.

Usage

```
vuong_structured(models, cpar.m1, copnames.m1, cpar.m2,
copnames.m2, y, grpsize)
```

Arguments

`models`	choose a number from (1,2,3,4). **1**: Model1 is bifactor, Model2 is bifactor. **2**: Model1 is second-order, Model2 is second-order. **3**: Model1 is second-order, Model2 is bifactor. **4**: Model1 is bifactor, Model2 is nested.
`cpar.m1`	vector of copula paramters for model 1, starting with copula parameters that link the items with common factor (bifactor), or group factors with common factor (second-order).
`cpar.m2`	vector of copula paramters for model 2, starting with copula parameters that link the items with common factor (bifactor), or group factors with common factor (second-order).
`copnames.m1`	vector of names of copula families for model 1, starting with copulas that link the items with common factor (bifactor), or group factors with common factor (second-order).
`copnames.m2`	vector of names of copula families for model 2, starting with copulas that link the items with common factor (bifactor), or group factors with common factor (second-order).
`y`	matrix of ordinal data.
`grpsize`	vector indicating the size for each group, e.g., c(4,4,4) indicating four items in all three groups.

Value

A vector containing the following components:

`z`	the test statistic.
`p.value`	the p-value.
`CI.left`	lower/left endpoint of 95% confidence interval.
`CI.right`	upper/right endpoint of 95% confidence interval.

Author(s)

Sayed H. Kadhem <s.kadhem@uea.ac.uk>
Aristidis K. Nikoloulopoulos <a.nikoloulopoulos@uea.ac.uk>

References

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman and Hall/CRC.

Kadhem, S.H. and Nikoloulopoulos, A.K. (2021) Bi-factor and second-order copula models for item response data. *Arxiv e-prints*, <arXiv:2102.10660>. https://arxiv.org/abs/2102.10660.

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

Examples

```
#-------------------------------------------------
# Setting quadreture points
nq <- 25
gl <- gauss.quad.prob(nq)
#-------------------------------------------------
#                     TAS Data
#-----------------             -----------------
data(TAS)
grp1=c(1,3,6,7,9,13,14)
grp2=c(2,4,11,12,17)
grp3=c(5,8,10,15,16,18,19,20)
ydat=TAS[,c(grp1,grp2,grp3)]

d=ncol(ydat);d
n=nrow(ydat);n

#Rearrange items within testlets
g1=length(grp1)
g2=length(grp2)
g3=length(grp3)

grpsize=c(g1,g2,g3)#group size
#number of groups
ngrp=length(grpsize)
#-------------------------------------------------
# M1 bifactor - M2 bifactor
cpar.m1 = rep(0.6,d*2)
copnames.m1 = rep("bvn",d*2)
cpar.m2 = rep(1.6,d*2)
copnames.m2 = rep("rgum",d*2)
vuong.bifactor = vuong_structured(models=1, cpar.m1, copnames.m1,
                 cpar.m2, copnames.m2,
                 y=ydat, grpsize)

# M1 seconod-order - M2 seconod-order
cpar.m1 = rep(0.6,d+ngrp)
copnames.m1 = rep("bvn",d+ngrp)
cpar.m2 = rep(1.6,d+ngrp)
copnames.m2 = rep("rgum",d+ngrp)
vuong.second_order = vuong_structured(models=2, cpar.m1,
```

```
    copnames.m1, cpar.m2, copnames.m2, y=ydat, grpsize)

# M1 seconod-order - M2 bifactor
cpar.m1 = rep(0.6,d+ngrp)
copnames.m1 = rep("bvn",d+ngrp)
cpar.m2 = rep(1.6,d*2)
copnames.m2 = rep("rgum",d*2)
vuong.2ndO_bif = vuong_structured(models=3, cpar.m1, copnames.m1,
                 cpar.m2, copnames.m2,
                 y=ydat, grpsize)

# M1 bifactor - M2 seconod-order
cpar.m1 = rep(0.6,d*2)
copnames.m1 = rep("bvn",d*2)
cpar.m2 = rep(1.6,d+ngrp)
copnames.m2 = rep("rgum",d+ngrp)
vuong.bif_2ndO = vuong_structured(models=4, cpar.m1, copnames.m1,
                 cpar.m2, copnames.m2,
                 y=ydat, grpsize)
```

Index