the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021)

published by John Wiley & Sons Ltd
on behalf of British Psychological Society

www.wileyonlinelibrary.com

Factor copula models for mixed data

Sayed H. Kadhem and Aristidis K. Nikoloulopoulos
School of Computing Sciences, University of East Anglia, Norwich, UK

We develop factor copula models to analyse the dependence among mixed continuous and discrete responses. Factor copula models are canonical vine copulas that involve both observed and latent variables, hence they allow tail, asymmetric and nonlinear dependence. They can be explained as conditional independence models with latent variables that do not necessarily have an additive latent structure. We focus on important issues of interest to the social data analyst, such as model selection and goodness of fit. Our general methodology is demonstrated with an extensive simulation study and illustrated by reanalysing three mixed response data sets. Our studies suggest that there can be a substantial improvement over the standard factor model for mixed data and make the argument for moving to factor copula models.

1. Introduction

It is very common in social science (e.g., in surveys) to deal with data sets that have mixed continuous and discrete responses. 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 framework.

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 are a manifestation of underlying continuous variables that usually follow the normal distribution (e.g., Lee, Poon, & Bentler, 1992; Muthén, 1984; Quinn, 2004); and the response function approach that postulates distributions on the observed variables conditional on the latent variables usually being from the exponential family (e.g., Huber, Ronchetti, & Victoria-Feser, 2004; Moustaki, 1996; Moustaki & Knott, 2000; Moustaki & Victoria-Feser, 2006; Wedel & Kamakura, 2001). 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 multivariate normal (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 & de Leeuw, 1987).

*Correspondence should be addressed to Aristidis K. Nikoloulopoulos, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK (email: a.nikoloulopoulos@uea.ac.uk).

DOI:10.1111/bmsp.12231

The underlying variable approach calls the MVN distribution as a latent model for the discrete responses, and therefore maximum likelihood (ML) estimation requires multidimensional integrations (Nikoloulopoulos, 2013, 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, for example, Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). The response function approach, with the dependence coming from p latent (unobservable) variables/factors, where $p \ll d$ (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 latent variables, which is not valid in the realistic scenario of tail asymmetry or tail dependence existing in the mixed data. Ma and Genton (2010), Montanari and Viroli (2010), and Irincheeva, Cantoni, and Genton (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, Cantoni, and Genton (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 (He, Li, Edmondson, Rader, & Li, 2012; Hoff, 2007; Jiryaie, Withanage, Wu, & de Leon, 2016; Shen & Weissfeld, 2006; Song, Li, & Yuan, 2009), 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, Hong, Czado, and Ghosh (2015), and Zilko and Kurowicka (2016) employed vine copulas to model mixed data. Vine copulas have two major advantages over the MVN copula, as emphasized in Panagiotelis, Czado, Joe, and Stöber (2017). The first is that the computational complexity of computing the joint probability distribution function grows 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 this paper 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). Factor copula models are more interpretable and fit better than vine copula models, when dependence can be explained through latent variables. Furthermore, they are closed under margins, that is, lower-order marginals belong to the same parametric family of copulas and a different permutation of the observed variables has exactly the

same distribution. This is not the case for vine copulas without latent variables, where a different permutation of the observed variables could lead to a different distribution.

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 & Joe, 2013; Nikoloulopoulos & Joe, 2015) was mainly based on simple diagnostics. In addition to simple diagnostics based on semi-correlations, we propose a heuristic method that automatically selects the bivariate parametric copula families. With regard to the issue of goodness-of-fit testing, we propose a technique based on the M_2 goodness-of-fit statistic (Maydeu-Olivares & 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 & Knott, 2000).

The remainder of the paper proceeds as follows. Section 2 introduces the factor copula models for mixed data and provides choices of parametric bivariate copulas with latent variables. Section 3 provides estimation techniques and computational details. Sections 4 and 5 propose methods for model selection and goodness of fit, respectively. Section 6 presents applications of our methodology to three mixed response data sets. Section 7 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 some discussion in Section 8, followed by a brief section with software details.

2. 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 $\mathbf{Y} = (Y_1, ..., Y_d)$ are conditionally independent given p latent variables $X_1, ..., X_p$. 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 one-factor and two-factor cases. 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 successfully performed as we strategically assume that the factors or latent variables are independent.

For the one-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 $\Pr(X_1 \leq x, Y_j \leq y) = C_{X_1 j}(x, F_j(y))$ for $0 \leq x \leq 1$ where F_j is the cumulative distribution function of Y_j. Then it follows that

$$F_{j|X_1}(y|x) := \Pr(Y_j \leq y | X_1 = x) = \frac{\partial C_{X_1 j}(x, F_j(y))}{\partial x}. \tag{1}$$

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

[1] We mean the density of $\mathbf{Y}$ with respect to 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 $\mathbb{R}$.

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^1 \prod_{j=1}^{d} f_{j|X_1}\left(y_j|x\right) \mathrm{d}x, \tag{2}$$

where

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

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 two-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 one-factor model, and let $C_{X_2 j}$ be a bivariate copula such that

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

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

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

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

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2 j|X_1}(x_2, y_j|x_1) \mathrm{d}x_1 \mathrm{d}x_2, \tag{3}$$

where

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

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 in the choices of the parametric marginal F_j or copula $\{C_{X_1 j}, C_{X_2 j}\}$ distributions.

2.1. Choices of bivariate copulas with latent variables

We provide choices of parametric bivariate copulas that can be used to link the latent to the observed variables. We will consider copula families that have different tail dependence (Joe, 1993) or tail order (Hua & Joe, 2011).

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 $0 \le u_1, u_2 \le 1$. 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 $(1-U_1,1-U_2)\sim \hat{C}$, where $\hat{C}(u_1,u_2)=u_1+u_2-1+C(1-u_1,1-u_2)$ 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. Under some regularity conditions (e.g., existing finite density in the interior of the unit square, ultimately monotone in the tail), if there exist $\kappa_L(C)>0$ and some $L(u)$ that is slowly varying at 0^+ (i.e., $\frac{L(ut)}{L(u)}\sim 1$, as $u\to 0^+$ for all $t>0$), 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), that is, $\bar{C}(1-u,1-u)\sim u^{\kappa_U(C)}L^*(u)$ as $u\to 0^+$, where $\bar{C}$ 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.

Having provided brief definitions of tail dependence and tail order, we provide below a list of bivariate parametric copulas with varying tail behaviour:

- 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;
- reflection symmetric copulas with tail quadrant independence ($\kappa_L=\kappa_U=2$), such as the Frank copula;
- 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 $\kappa_U=1$;
- reflection symmetric copulas with tail dependence, such as the t_ν copula with ν the degrees of freedom and $\kappa_L=\kappa_U=1$;
- 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 $\kappa_U=1$;
- reflection asymmetric copulas with tail quadrant independence, such as the BB8 and BB10 copulas.

The BVN, Frank, and t_ν are comprehensive copulas, that is, they interpolate between countermonotonicity (perfect negative dependence) and comonotonicity (perfect positive dependence). The other aforementioned parametric families of copulas (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

- $(1-U_1,U_2)\sim \hat{C}^{(1)}$, where $\hat{C}^{(1)}(u_1,u_2)=u_2-C(1-u_1,u_2)$ is the 1-reflected copula of C with negative lower-upper tail dependence;
- $(U_1,1-U_2)\sim \hat{C}^{(2)}$, where $\hat{C}^{(2)}(u_1,u_2)=u_1-C(u_1,1-u_2)$ 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, to depict the concepts of refection symmetric and asymmetric tail dependence and quadrant tail independence, we show contour plots of the corresponding copula densities with standard normal margins and dependence parameters corresponding to a Kendall's τ value of .5. Sharper corners (relative to ellipse) indicate tail dependence.

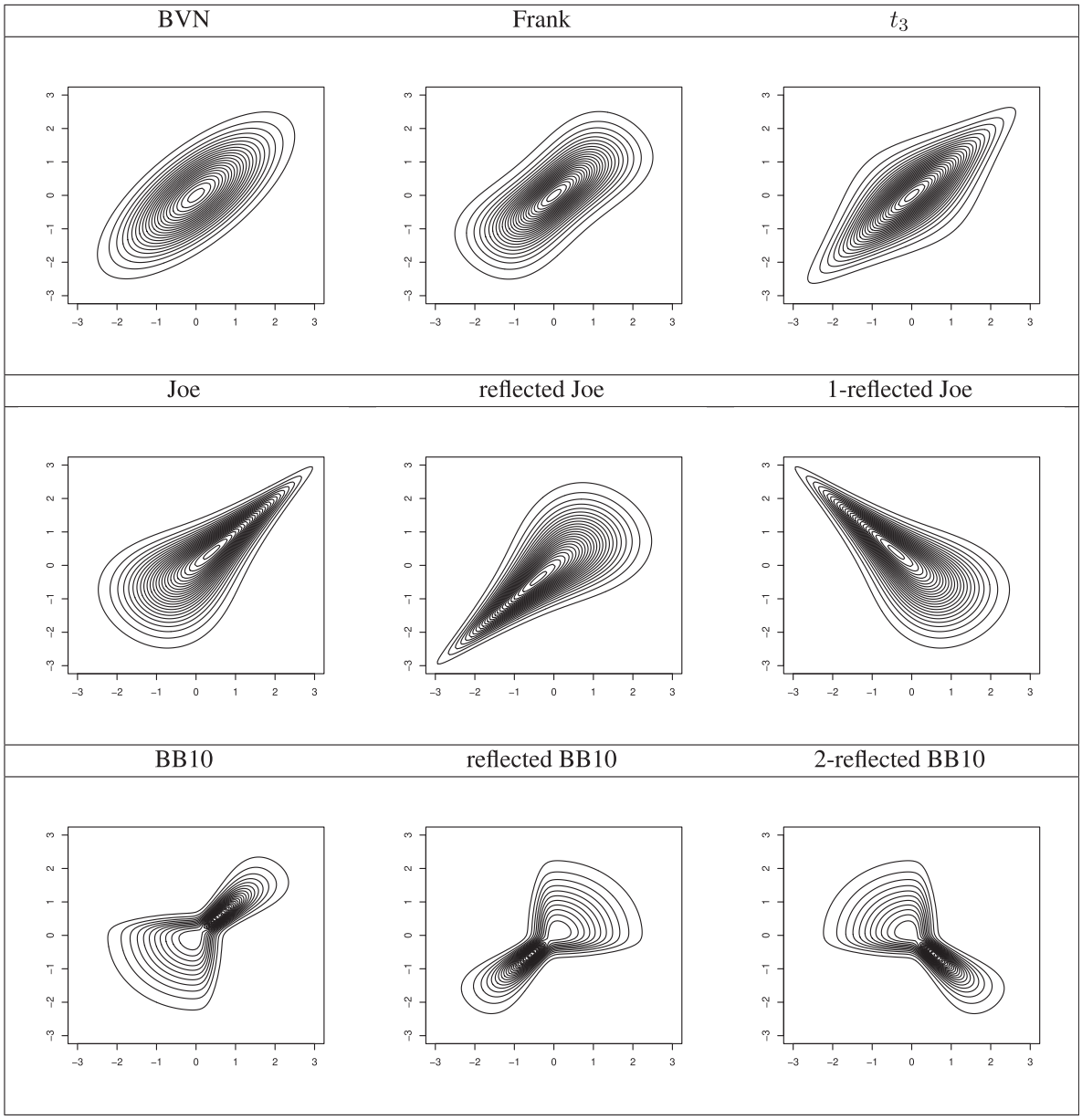

Figure 1. Contour plots of bivariate copulas with standard normal margins and dependence parameters corresponding to a Kendall's τvalue of .5 in absolute value.

2.2. 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 joint upper or lower tail probability 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, Joe, & Li, 2012). The correlations of normal scores in the upper and lower tail (hereafter semi-correlations) are defined as (Joe, 2014, p. 71):

$$\begin{aligned}\rho_N^+ &= \mathrm{Cor}(Z_{j_1}, Z_{j_2} | Z_{j_1} > 0, Z_{j_2} > 0) \\ &= \frac{\int_0^\infty \int_0^\infty z_1 z_2 \phi(z_1)\phi(z_2) c(\Phi(z_1), \Phi(z_2)) \mathrm{d}z_1 \mathrm{d}z_2 - \left(\int_0^\infty z\phi(z)(1 - C_{2|1}(0.5|\Phi(z)))\mathrm{d}z\right)^2 / C(0.5, 0.5)}{\int_0^\infty z^2 \phi(z)(1 - C_{2|1}(0.5|\Phi(z)))\mathrm{d}z - \left(\int_0^\infty z\phi(z)(1 - C_{2|1}(0.5|\Phi(z)))\mathrm{d}z\right)^2 / C(0.5, 0.5)},\end{aligned}$$

$$\begin{aligned}\rho_N^- &= \mathrm{Cor}(Z_{j_1}, Z_{j_2} | Z_{j_1} < 0, Z_{j_2} < 0) \\ &= \frac{\int_{-\infty}^0 \int_{-\infty}^0 z_1 z_2 \phi(z_1)\phi(z_2) c(\Phi(z_1), \Phi(z_2)) \mathrm{d}z_1 \mathrm{d}z_2 - \left(\int_{-\infty}^0 z\phi(z) C_{2|1}(0.5|\Phi(z))\mathrm{d}z\right)^2 / C(0.5, 0.5)}{\int_{-\infty}^0 z^2 \phi(z) C_{2|1}(0.5|\Phi(z))\mathrm{d}z - \left(\int_{-\infty}^0 z\phi(z) C_{2|1}(0.5|\Phi(z))\mathrm{d}z\right)^2 / C(0.5, 0.5)},\end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ is the univariate normal cdf and density, respectively.

Note in passing that for the BVN copula $\rho_N^+ = \rho_N^-$ and that it has a closed form (see Joe, 2014, p. 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 1 has semi-correlations for all the aforementioned bivariate parametric copulas with $\tau = \{.3, .5, .7\}$. 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

$$Y_j = y_j \quad \text{if } \alpha_{y_j - 1, j} \le Z_j \le \alpha_{y_j j}, y_j = 1, \ldots, K_j, \tag{4}$$

where K_j is the number of categories of Y_j and $\alpha_{1j}, \ldots, \alpha_{K_j - 1, j}$ are the univariate cutpoints (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 variable 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

$$\hat{\rho}_N = \arg\max_\rho \sum_{i=1}^n \log\left(\Phi_2(\alpha_{y_{i1}}, \alpha_{y_{i2}}; \rho) - \Phi_2(\alpha_{y_{i1}-1}, \alpha_{y_{i2}}; \rho) - \Phi_2(\alpha_{y_{i1}}, \alpha_{y_{i2}-1}; \rho) + \Phi_2(\alpha_{y_{i1}-1}, \alpha_{y_{i2}-1}; \rho)\right)$$

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

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

with $z_{ij} = \Phi\left((n+1)^{-1} \sum_{i=1}^n \mathbf{1}(Y_{ij} \le y_{ij})\right)$, respectively.

3. Estimation

We use a two-stage copula modelling approach to the estimation of a multivariate model that borrows the strengths of the semi-parametric and inference function for margins

Table 1. Lower semi-correlations ρ_N^-, upper semi-correlations ρ_N^+, lower tail dependence λ_L, and upper tail dependence λ_U, with $\tau = \{.3, .5, .7\}$ for one-parameter and two-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

(IFM) approach in Genest, Ghoudi, and Rivest (1995) and Joe (2005), respectively. Suppose that the data are $y_{ij}, j = 1, ..., d, i = 1, ..., n$, where i is an index for individuals or clusters and j is an index for the within-cluster measurements. For $i = 1, ..., n$, we start from a d-variate sample $y_{i1}, ..., y_{id}$ from which d estimators $F_1(y_{i1}), ..., F_d(y_{id})$ can be obtained. We use these to transform the $y_{i1}, ..., y_{id}$ sample into a uniform sample $u_{i1} = F_1(y_{i1}), ..., u_{id} = F_d(y_{id})$ on $[0, 1]^d$ and then fit the factor copula model at the second step. For continuous and discrete data y_{ij}, we use nonparametric and parametric univariate distributions, respectively, to transform the data y_{ij} into copula data $u_{ij} = F_j(y_{ij})$, that is, 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.

3.1. Univariate modelling

For continuous random variables, we estimate each marginal distribution nonparametrically by the empirical distribution function of Y_j, namely,

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

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 & 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 Equation (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 $\boldsymbol{\gamma}_j = (\alpha_{1j}, ..., \alpha_{K_j-1,j})$ is the vector of the univariate cutpoints.
- For a count response variable Y_j, we use the negative binomial distribution (Lawless, 1987). This allows for over-dispersion and its probability mass function is

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

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 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 $\boldsymbol{\gamma}_j$. This is equivalent to 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 demonstrated in Kim, Silvapulle, and Silvapulle (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.

3.2. Copula modelling

Having estimated the univariate marginal distributions, we proceed to estimation of the dependence parameters. For the one-factor and two-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 $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_j, \theta_j : j = 1, ..., d\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_j, \theta_j, \delta_j : j = 1, ..., d\}$ denote the set of all parameters for

the one- and two-factor model, respectively. Estimation can be achieved by maximizing the joint log-likelihood

$$\ell_{\mathbf{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\mathbf{Y}}(y_{i1},...,y_{id};\boldsymbol{\theta}) \tag{5}$$

over the copula parameters θ_j or $\delta_j, j = 1,...,d$, 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 (the logarithm of the joint likelihood), while the gradients are computed numerically and the Hessian matrix of the second-order derivatives is updated at each iteration. The standard errors (*SE*s) of the estimates can be obtained via the gradients and the Hessian computed numerically during the maximization process. These *SE*s are adequate to assess the flatness of the log-likelihood. Proper *SE*s that account for the estimation of univariate parameters can be obtained by maximizing the joint likelihood in equation (5) in 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 & Secrest, 1966). To compute one-dimensional integrals for the one-factor model, we use the approximation

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

where $\{x_q : q = 1,...,n_q\}$ are the quadrature points and $\{w_q : q = 1,...,n_q\}$ are the quadrature weights. To compute two-dimensional integrals for the two-factor model, the approximation uses Gauss–Legendre quadrature points in a double sum:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \int_0^1 \int_0^1 \prod_{j=1}^{d} f_{X_2 j|X_1}(\mathrm{x}_2, y_j|x_1)\,\mathrm{d}x_1\mathrm{d}x_2 \\ &\approx \sum_{q_1=1}^{n_q} \sum_{q_2=1}^{n_q} w_{q_1} w_{q_2} \prod_{j=1}^{d} f_{X_2 j|X_1}(x_{q_2}, y_j|x_{q_1}). \end{aligned}$$

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 $n_q = 25$ is adequate with good precision.

4. 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 of the tail asymmetry based on semi-correlations is not consistent or clear. For multivariate mixed data, it is not feasible 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

for 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 pairs where necessary to account for tail asymmetry (discrete data) or tail dependence (continuous data).

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

1. For $j = 1, ..., d$, estimate the marginal distributions $F_j(y)$.
2. Fit the one-factor copula model with Frank copulas to link each of the d observed variables with the latent variable, that is, maximize the log-likelihood function of the factor copula model in Equation (5) over the vector of copula parameters $(\theta_1, ..., \theta_d)$.
3. If the jth linking copula has $\hat{\theta}_j > 0$, then select a set of copula candidates with the ability to interpolate between independence and comonotonicity, otherwise select a set of copula candidates with ability to interpolate between countermonotonicity and independence.
4. For $j = 1, ..., d$:
   a. fit all the possible one-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.

5. 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 Vuong's test (Vuong, 1989) to show if a factor copula model provides a better fit than the standard factor model with a latent additive structure, that is a factor copula model with BVN bivariate linking copulas (Krupskii & Joe, 2013; Nikoloulopoulos & Joe, 2015). 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, Czado, & Aas, 2012; Joe, 2014; Nikoloulopoulos, 2017). We provide specific details in Section 5.1.

Furthermore, to assess the overall goodness of fit of the factor copula models for mixed data, we will make appropriate use of the limited information M_2 statistic (Maydeu-Olivares & 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 5.2.

5.1. 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 $f_{\mathbf{Y}}^{(1)}$ and $f_{\mathbf{Y}}^{(2)}$, respectively. We can compare

$$\Delta_{1f_{\mathbf{Y}}} = n^{-1}\left[\sum_{i=1}^{n}\left\{E_{f_{\mathbf{Y}}}\log f_{\mathbf{Y}}(\mathbf{y}_i) - E_{f_{\mathbf{Y}}}\log f_{\mathbf{Y}}^{(1)}(\mathbf{y}_i;\boldsymbol{\theta}_1)\right\}\right],$$

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

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 ML estimates based on models 1 and 2, respectively.

Model 1 is closer to the true $f_{\mathbf{Y}}$, i.e., it 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 ML estimates $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ is

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

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

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

where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$. Hence, its 95% confidence interval (CI) is $\overline{D} \pm 1.96 \times \frac{1}{\sqrt{n}}\sigma$.

5.2. M_2 goodness-of-fit statistic

Since the M_2 statistic has been developed for multivariate ordinal data (Maydeu-Olivares & Joe, 2006), we propose to first transform the continuous and count variables to ordinal and then calculate the M_2 statistic at the ML 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 into 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 by 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 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 for categorizing 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 $Y_1,...,Y_d$ (both the original and the transformed ones) where the jth $(1 \le j \le d)$ variable consists of $K_j \ge 2$ categories labelled $0,1,...,K_j-1$. Consider the set of univariate and bivariate residuals that do not include category 0. This is a residual vector of dimension

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

For a factor copula model with parameter vector $\boldsymbol{\theta}$ of dimension q, let $\boldsymbol{\pi}_2(\boldsymbol{\theta}) = (\dot{\pi}_1(\boldsymbol{\theta})^T, \dot{\pi}_2(\boldsymbol{\theta})^T)^T$ 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 $\mathbf{p}_2 = (\dot{\mathbf{p}}_1^T, \dot{\mathbf{p}}_2^T)^T$ 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

$$M_2 = M_2(\hat{\boldsymbol{\theta}}) = n(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}))^T \mathbf{C}_2(\hat{\boldsymbol{\theta}})(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})), \tag{6}$$

with

$$\mathbf{C}_2(\boldsymbol{\theta}) = \boldsymbol{\Xi}_2^{-1} - \boldsymbol{\Xi}_2^{-1}\Delta_2(\Delta_2^T\boldsymbol{\Xi}_2^{-1}\Delta_2)^{-1}\Delta_2^{\mathrm{T}}\boldsymbol{\Xi}_2^{-1} = \Delta_2^{(c)}\left([\Delta_2^{(c)}]^T\boldsymbol{\Xi}_2\Delta_2^{(c)}\right)^{-1}[\Delta_2^{(c)}]^T, \tag{7}$$

where $\Delta_2 = \partial\boldsymbol{\pi}_2(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}$ is an $s \times q$ matrix with the derivatives of all the univariate and bivariate marginal probabilities with respect to the model parameters, $\Delta_2^{(c)}$ is an $s \times (s-q)$ orthogonal complement to Δ_2 such that $[\Delta_2^{(c)}]^T\Delta_2 = \mathbf{0}$, and $\boldsymbol{\Xi}_2 = \mathrm{diag}(\boldsymbol{\pi}_2(\boldsymbol{\theta})) - \boldsymbol{\pi}_2(\boldsymbol{\theta})\boldsymbol{\pi}_2(\boldsymbol{\theta})^T$ is the $s \times s$ covariance matrix of all the univariate and bivariate marginal sample proportions, excluding category 0. Due to equality in (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 Δ_2 for factor copula models we refer the interested reader to Nikoloulopoulos and Joe (2015).

6. Applications

In this section we illustrate the proposed methodology by reanalysing three mixed response data sets.

Initially, we use the diagnostic method in Joe (2014, pp. 245–246) to show that each data set (or, more precisely, the correlation matrix of the observed variables for each data set) 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 $D_3 = \log(\det(\mathbf{R}_{\text{model}})) - \log(\det(\mathbf{R}_{\text{observed}})) + \text{tr}(\mathbf{R}_{\text{model}}^{-1}\mathbf{R}_{\text{observed}}) - d$.

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 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.

Having discussed 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 copula model, to capture any type of reflection asymmetric dependence, by using the proposed algorithm in Section 4. 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 two-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 the other $2d-1$ parameters. We report the varimax transform (Kaiser, 1958) of the loadings (a reparametrization of the $2d$ parameters), converted to factor copula parameters via the relations

$$\theta_j = \beta_{j1}, \quad \delta_j = \frac{\beta_{j2}}{(1-\beta_{j1}^2)^{1/2}}, \tag{8}$$

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

If the number of parameters is not the same between the models, we use the AIC as a rough diagnostic measure of goodness of fit between the models, otherwise we use the likelihood at the ML estimates. We further compute Vuong's tests with model 1 as the factor copula model with BVN copulas (i.e., 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 τs in $(-1,1)$ via the relations in Joe (2014, Chapter 4); *SE*s are also converted via the delta method. For the model that provides the best fit, we provide the estimates and *SE*s that are obtained by maximizing the joint likelihood in equation (5) in one step over $\boldsymbol{\theta}$. Although the two-stage estimation approach in Section 3 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 Δ_2 having one less column .

6.1. Political-economic data set

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 its constituents' advantage (see, for example, North & Weingast, 1989). Quinn (2004) used five mixed variables, namely, the black-market premium in each country (continuous, used as a proxy for illegal economic activity), productivity as measured by real gross domestic product per worker at 1985 international prices (continuous), the independence of the national judiciary (binary; 1 if the judiciary is judged to be independent and 0 otherwise), and two ordinal variables measuring the lack of expropriation risk and lack of corruption. The data set and a complete description thereof can be found in Quinn (2004) or in the `R` package `MCMCpack` (Martin, Quinn, & Park, 2011). Note that since the black-market premium is negatively associated with the remaining variables (from the context), we reorient it, leading to positive dependence among all the observed variables.

Table 2 shows that the sample correlation matrix of the mixed responses has a one-factor structure based on linear factor analysis (large D_3 is due to the small sample size as demonstrated using simulated data in Section 7). The sample semi-correlations in Table 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 3 gives the estimated parameters, their *SE*s on Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's

Table 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 one and two factors for the political-economic risk data

Pairs of variables		ρ_N	ρ_N^-	ρ_N^+
BM	GDP	.53	−.04	.57
BM	IJ	.61	–	–
BM	XPR	.67	.88	.63
BM	CRP	.62	.16	.55
GDP	IJ	.78	–	–
GDP	XPR	.55	.11	.75
GDP	CRP	.77	.24	.63
IJ	XPR	.91	–	–
IJ	CRP	.87	–	–
XPR	CRP	.76	.71	.71

No. of factors	D_1	D_2	D_3
1	0.16	0.04	0.91
2	0.06	0.01	0.22

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

Table 3. Estimated parameters, their standard errors (*SE*) on 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

	BVN[a]		t_5		Frank		Selected model		
One-factor	$\hat{\tau}$	*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	<.01		<.01		<.01		.60		

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

tests, and the M_2 statistics for the one-factor copula models. Table 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.

In all the fitted models the estimated Kendall's τs are similar. Kendall's τ only accounts for the dependence dominated by the middle of the data, and it is expected to be similar among different families of copulas. However, the tail dependence and tail order vary, as explained in Section 2.1, and they are properties to consider when choosing among different families of copulas (Nikoloulopoulos & 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 five categories using the unsupervised strategy in Section 5.2; 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 political-economic certainty.

6.2. General Social Survey

Hoff (2007) analysed seven demographic variables for 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 five 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, supplementary materials).

Table 4 shows that the sample correlation matrix of the mixed responses has a two- or even three-factor structure based on linear factor analysis. The direction of the tail asymmetry based on sample semi-correlations in Table 4 is not consistent, and this shows the usefulness of the proposed model selection technique. Table 5 gives the estimated parameters, their *SE*s on Kendall's τ scale, the joint log-likelihoods, the 95% CIs of Vuong's tests, and the M_2 statistics for the one-factor and two-factor copula models. The best fit for the one-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 four (18–24, 25–44, 45–64, and 65+) and five (0–10, 11–19, 20–29, 30–40, and 41+)

Table 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 data set

Pairs of variables		ρ_N	ρ_N^-	ρ_N^+
Income	Age	.29	.48	.23
Income	Degree	.52	.24	.33
Income	Pincome	.14	.02	.28
Income	Pdegree	.24	.04	.08
Income	Child	.22	.23	.01
Income	Pchild	−.09	.06	.00
Age	Degree	.06	.22	−.04
Age	Pincome	−.11	−.02	.12
Age	Pdegree	−.14	−.42	.44
Age	Child	.58	.36	.26
Age	Pchild	.12	.18	.07
Degree	Pincome	.21	.17	−.05
Degree	Pdegree	.46	.46	.41
Degree	Child	−.11	−.10	−.09
Degree	Pchild	−.25	−.14	−.30
Pincome	Pdegree	.44	.44	.34
Pincome	Child	−.16	−.15	.11
Pincome	Pchild	−.23	.13	−.30
Pdegree	Child	−.21	.08	.10
Pdegree	Pchild	−.34	.19	−.32
Child	Pchild	.20	−.11	−.06

No. of 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 5. Estimated parameters, their standard errors (*SE*) on Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's statistics, and the M_2 statistics for the one- and two-factor copula models for the General Social Survey data set

	BVN[a]		t_9		Frank		Selected model		
One-factor	$\hat{\tau}$	*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	2rJoe	−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	rGumbel	0.56	0.05
Child	−0.20	0.04	−0.19	0.04	−0.19	0.04	2rJoe	−0.14	0.03
Pchild	−0.32	0.03	−0.31	0.04	−0.32	0.04	2rGumbel	−0.27	0.03
ℓ	−3,425.39		−3,420.56		−3,433.83		−3,397.79		
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	<.001		<.001		<.001		<.001		

	BVN[a]	t_9	Frank		Selected model		
Two-factor	$\hat{\tau}$	$\hat{\tau}$	$\hat{\tau}$	*SE*	Copulas	$\hat{\tau}$	*SE*
First factor							
Income	0.36	0.35	0.13	0.04	rGumbel	0.34	0.03
Age	−0.05	−0.06	0.50	0.05	rJoe	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	1rJoe	−0.13	0.04
Pdegree	0.48	0.50	−0.31	0.05	1rJoe	−0.13	0.04
Child	−0.13	−0.14	0.52	0.05	rJoe	0.44	0.04
Pchild	−0.28	−0.28	0.23	0.04	Gumbel	0.11	0.03
Second factor							
Income	0.38	0.41	0.50	0.06	Gumbel	0.40	0.04
Age	0.54	0.55	0.21	0.04	2rJoe	−0.14	0.03
Degree	0.14	0.17	0.57	0.07	rJoe	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	2rGumbel	−0.26	0.03
ℓ	−3,286.80	−3,278.88	−3,300.07		−3,235.86		
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	<.001	<.001	<.001		0.13		

Note. [a]The resulting model is the same as the standard factor model. pdemographic: demographic variable of respondent's parents. rCopula: reflected copula; 1rCopula: 1-reflected copula; 2rCopula: 2-reflected copula.

categories, respectively, and the numbers of children of the survey respondent and respondent's parents were treated as ordinal where the fourth (more than 3 children) and eighth (more than 7 children) category, respectively, contained all the high counts.

The two-factor copula models with BVN, t_ν, and Frank copulas provide some improvement over the one-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 *SE*s) 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 (8).

The selected two-factor copula model using the algorithm in Section 4 shows improvement over the standard factor model according to Vuong's statistic and better fit according to the M_2 statistic; it changes a p-value less than .001 to one greater than .10. For the two-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 $\hat{\theta}_j$ and $\hat{\delta}_j$ and then to loadings using the relations in (8)) show that one factor is loaded only on the demographic variables of the respondent's parents.

6.3. 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 $n = 9,960$ 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 of the household's possession).

With simple descriptive statistics such as scatter plots of the original data, Irincheeva et al. (2012b), 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 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 exist on the data. For all the pairs the upper semi-correlation is larger, and interestingly, contrasting the bivariate normal scores plots in Figure 2 with the contour plots in Figure 1, it is apparent that for the continuous variables the linking copulas might be the BB10 copulas.

Table 6 shows that the sample correlation matrix of the mixed responses has a two-factor structure based on linear factor analysis. The sample semi-correlations in Table 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 7 gives the estimated parameters, their *SE*s on Kendall's tau scale, the joint log-likelihoods, the 95% CIs of Vuong's tests, and the M_2 statistics for the one-factor and two-factor copula models. The best-fitting one- and two-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 among the mixed responses based on the M_2 statistic (Table 7,

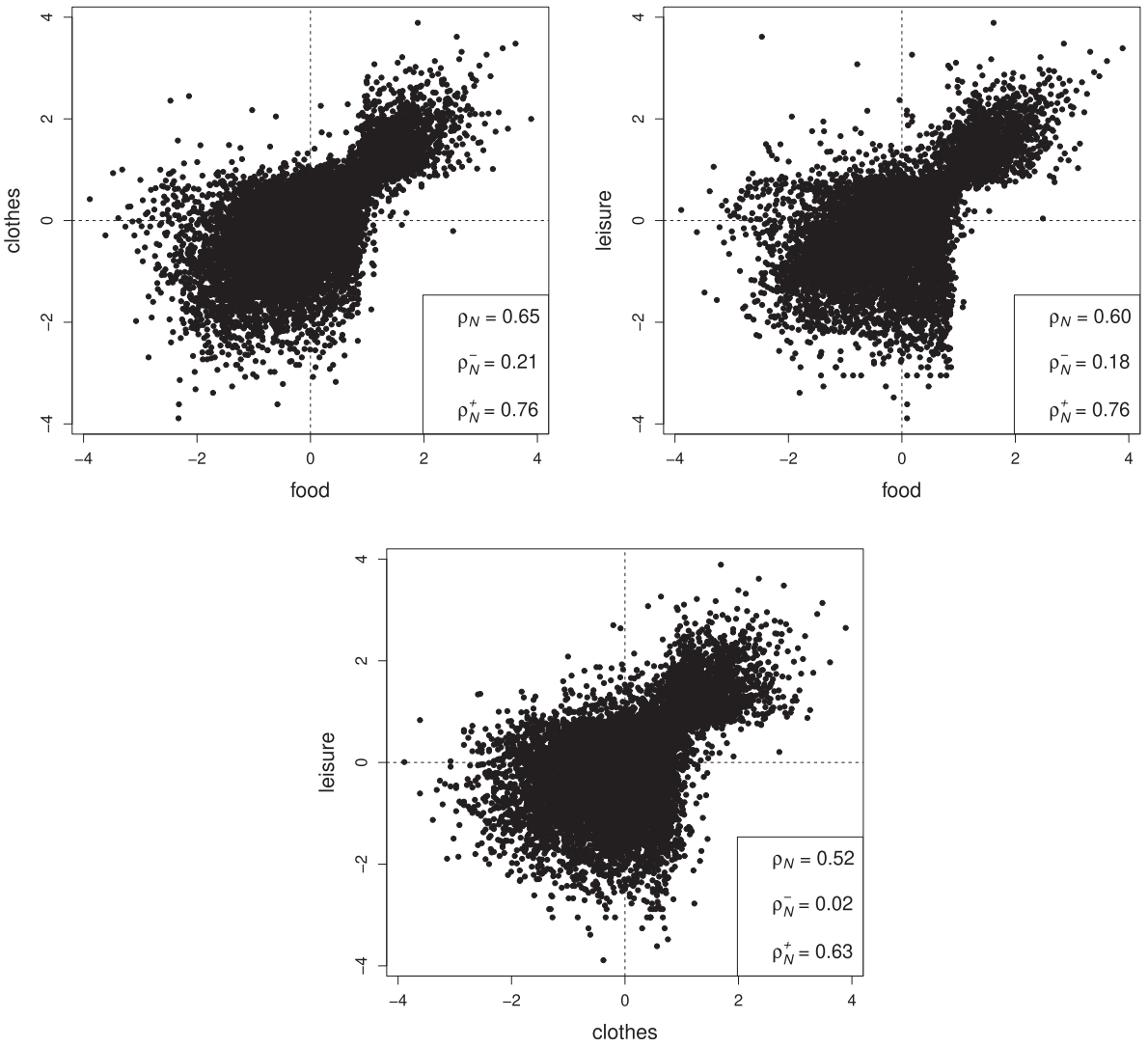

Figure 2. Bivariate normal scores plots, along with correlations and semi-correlations for the continuous data from the Swiss Consumption Survey.

one factor). To apply the M_2 statistic, we transformed the continuous variables to ordinal with three categories using the unsupervised strategy in Section 5, and the count variable bicycle was treated as ordinal where the sixth category contained all the high counts (five bicycles or more).

While it is revealed that the selected two-factor copula model is the best model (lowest AIC) and there is substantial improvement over the standard two-factor model, it is not apparent from the M_2 statistic that the response patterns are satisfactorily explained by even two 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 & Joe, 2014). In Table 8 we list the maximum deviations of observed and expected counts for each bivariate margin, that is, $D_{j_1 j_2} = n \max |p_{j_1,j_2,y_1,y_2} - \pi_{j_1,j_2,y_1,y_2}(\hat{\boldsymbol{\theta}})|$. 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 509/9,960 maximum occurs in the BVN factor copula model, while this drops to 133/9,960 in the selected two-factor copula model.

Table 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 data set

Pairs of variables		ρ_N	ρ_N^-	ρ_N^+
Food	Clothes	.65	.21	.76
Food	Leisure	.60	.18	.76
Food	Dishwasher	.31	–	–
Food	Car	.38	–	–
Food	Motorcycle	.11	–	–
Food	Bicycles	.21	.22	.02
Clothes	Leisure	.52	.02	.63
Clothes	Dishwasher	.23	–	–
Clothes	Car	.25	–	–
Clothes	Motorcycle	.07	–	–
Clothes	Bicycles	.18	.15	.02
Leisure	Dishwasher	.24	–	–
Leisure	Car	.18	–	–
Leisure	Motorcycle	.01	–	–
Leisure	Bicycles	.08	.04	.08
Dishwasher	Car	.43	–	–
Dishwasher	Motorcycle	.03	–	–
Dishwasher	Bicycles	.24	–	–
Car	Motorcycle	.18	–	–
Car	Bicycles	.26	–	–
Motorcycle	Bicycles	.21	–	–

No. of 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

For the selected two-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 shows that the one latent variable which is only associated with the continuous variables measures expenses, while the other which is associated with all the mixed variables measures possession.

7. Simulations

An extensive simulation study was 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-

Table 7. Estimated parameters, their standard errors (*SE*) on Kendall's τ scale, joint log-likelihoods, the 95% CIs of Vuong's statistics, and the M_2 statistics for the one- and two-factor copula models for the Swiss Consumption Survey data set

	BVN[a]		t_5		Frank		Selected model		
One-factor	$\hat{\tau}$	*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	55,004.24		54,221.36		55,105.88		48,932.32		
Vuong 95% CI			(0.032, 0.046)		(−0.015, 0.005)		(0.286, 0.324)		
M_2	2,775.73		2,734.05		2,808.53		1,626.54		
df	71		71		71		68		
p-value	<.001		<.001		<.001		<.001		
	BVN[a]		t_7		Frank		Selected model		
Two-factor	$\hat{\tau}$		$\hat{\tau}$	*SE*	$\hat{\tau}$	*SE*	Copulas	$\hat{\tau}$	*SE*
First 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
Second 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	54,245.91		53,482.23		53,514.75		46,233.00		
Vuong 95% CI			(0.032, 0.045)		(0.028, 0.046)		(0.386, 0.419)		
M_2	1,920.27		1,886.66		1,945.07		450.32		
df	65		64		64		59		
p-value	<.001		<.001		<.001		<.001		

Note. [a]The resulting model is the same as the standard factor model.

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 of the selected one- and two-factor copula models in the three application examples in Section 6. We set

Table 8. Maximum deviations D_{j_1,j_2} of observed and expected counts for each bivariate margin (j_1, j_2) for the one- and two-factor copula models for the Swiss Consumption Survey data set

	One-factor model				Two-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

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. Binary variables do not have tail asymmetries, hence parametric copulas are less distinguishable. Therefore instead of binary, we simulated from ordinal variables with three equally weighted categories.

Table 9 contains the simulated means and standard deviations (*SD*s) of the discrepancy measures D_1, D_2 and D_3. The resulting 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 performs well 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 successfully 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 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 ($\tau = .3$), moderate ($\tau = .5$) and strong dependence ($\tau = .7$). Representative results are shown in Table 10 for the Gumbel copula. Note that the count variable was treated as ordinal with five categories, where the fifth category contained all the counts greater than 3. The resulting biases, root mean square errors (RMSEs), and *SD*s, 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 is used in computing sample semi-correlations, their variability is larger

Table 9. Small sample of size $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected factor copula models in Section 6 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 resulting means and standard deviations (*SD*)

		D_1		D_2		D_3	
n	No. of factors	Mean	*SD*	Mean	*SD*	Mean	*SD*
Political-economic data set – one-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 – one-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 – one-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 – two-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 – two-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

Continued

Table 9. (*Continued*)

		D_1		D_2		D_3	
n	No. of factors	Mean	*SD*	Mean	*SD*	Mean	*SD*
	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

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 11 contains the resulting biases, RMSEs, and *SD*s, scaled by n, for the estimates obtained using the estimation approach in Section 3. The results show that the proposed estimation approach is highly efficient according to the simulated biases, *SD*s, and RMSEs. We further investigated the misspecification of the bivariate pair copulas by deriving the same statistics but from the one-factor model with BVN pair copulas (i.e., the standard one-factor model). Once again, the simulated data are based on the selected one-factor copula models in Section 6. Table 12 contains the resulting biases, RMSEs, and *SD*s, 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 Figure 1) are used to specify the true factor copula model (Table 12, Swiss Consumption Survey). As we have already mentioned, the Kendall's τ only accounts for 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 12, political-economic data set and General Social Survey).

Table 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 5.2. We also transformed the count variables to ordinal with $K = 5$ categories by treating them as ordinal, where the fifth 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 14 presents the number of times the true bivariate parametric copulas were 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 it is then difficult to distinguish among parametric families of copulas (Nikoloulopoulos & Karlis, 2008). For example,

- in the results from the two-factor model for the General Social Survey, the true copula for the first continuous variable (first factor) is the reflected Gumbel with $\tau = .34$ and is only selected a very small number of times. The algorithm instead selected with a high probability the reflected Joe (results not shown here due to space constraints), because both reflected Joe and Gumbel copulas provide similar dependence properties, i.e., lower tail dependence.

Table 10. Small sample of size $n = \{100, 300, 500\}$ simulations (10^4 replications) from the Gumbel copula with Kendall's $\tau = \{.3, .5, .7\}$ for mixed continuous, ordinal, and count data with resulting 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	.3	True values	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.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
	.5	True values	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.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
	.7	True values	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.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	.3	True values	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.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	−.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
	.5	True values	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.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
		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
	.7	True values	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.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

Continued

Table 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^+
500	.3	True values	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.46	.46	.16	.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
	.5	True values	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.67	.70	.36	.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
	.7	True values	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.85	.88	.64	.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 11. Small sample of size $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected factor copula models in Section 6 with resulting biases, root mean square errors (RMSE) and standard deviations (*SD*), scaled by n, for the estimated parameters

Political-economic data set – one-factor model

τ	.51			.58			.80			.68			.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 – one-factor model

τ	.30			−.14			.46			.33			.55			−.14			−.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 – one-factor model

τ	.69			.38			.39			.28			.23			.13			.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 11 (*Continued*)

General Social Survey – two-factor model	$n = 500$													
	First factor							Second factor						
τ	.34	.49	.18	−.13	−.13	.44	.11	.40	−.14	.65	.29	.49	−.24	−.26
*n*Bias	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
*n*SD	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
*n*RMSE	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 – two-factor model	$n = 500$													
	First factor							Second factor						
τ	.34	.36	.38	.19	.09	.02	.04	.53	.28	.30	.42	.35	.17	.27
*n*Bias	−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
*n*SD	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
*n*RMSE	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 12. Small sample of size $n = \{100, 300, 500\}$ simulations (10^4 replications) from the selected one-factor copula models in Section 6 with resulting biases, root mean square errors (RMSE) and standard deviations (*SD*), scaled by n, for the estimated parameters under an one-factor copula model with BVN copulas, i.e. the standard factor model

Political-economic data set – one-factor model

τ	.51			.58			.80			.68			.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 – one-factor model

τ	.30			−.14			.46			.33			.55			-.14			−.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 – one-factor model

τ	.69			.38			.39			.28			.23			.13			.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 13. Small sample of size $n = \{100, 300, 500\}$ distributions for M_2 (10^4 replications). Empirical rejection levels at $\alpha = \{.20, .10, .05, .01\}$, 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 5.2. Count variables are also transformed to ordinal with $K = 5$ categories by treating them as ordinal, where the fifth 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 data set – one-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
$\alpha = .20$	.183	.192	.197	.196	.194	.195	.196	.189	.190
$\alpha = .10$	.121	.125	.134	.122	.121	.119	.114	.109	.109
$\alpha = .05$	.083	.089	.098	.076	.077	.077	.072	.070	.067
$\alpha = .01$	.044	.046	.055	.036	.034	.037	.027	.030	.026
General Social Survey – one-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
$\alpha = .20$	.213	.220	.240	.202	.216	.203	.211	.228	.212
$\alpha = .10$	.110	.121	.122	.106	.118	.102	.118	.127	.108
$\alpha = .05$	.058	.070	.061	.054	.067	.051	.065	.073	.056
$\alpha = .01$	.013	.018	.014	.014	.019	.012	.016	.023	.011
Swiss Consumption Survey – one-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
$\alpha = .20$	.229	.239	.254	.214	.209	.221	.210	.202	.207
$\alpha = .10$	.121	.135	.147	.111	.104	.113	.105	.099	.103
$\alpha = .05$	.067	.076	.086	.056	.055	.060	.051	.053	.053
$\alpha = .01$	.016	.024	.030	.011	.013	.013	.012	.011	.012

	General Social Survey – two-factor model			Swiss Consumption Survey – two-factor model		
	$n = 500$			$n = 500$		
	$K = 3$	$K = 4$	$K = 5$	$K = 3$	$K = 4$	$K = 5$
df	154	232	322	65	119	185
Mean	154.8	234.0	323.3	65.6	119.7	185.5
$\alpha = .20$	.217	.234	.214	.217	.215	.217
$\alpha = .10$	.113	.131	.116	.114	.111	.113
$\alpha = .05$	.065	.075	.059	.060	.057	.060
$\alpha = .01$	.018	.022	.018	.013	.013	.017

- in the results from the two-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 parameters are close to 0 (independence).

Table 14. Frequencies of the true bivariate copula identified using the model selection algorithm from 100 simulation runs.

Political-economic data set – one-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 – one-factor model

	Continuous		Ordinal			Count	
n	Joe	2rJoe	t_5	t_5	rGumbel	2rJoe	2rGumbel
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 – one-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 – two-factor model

	Continuous		Ordinal			Count	
1st factor *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 *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

Continued

Table 14 (*Continued*)

Swiss Consumption Survey – two-factor model							
	Continuous			Ordinal			Count
1st factor	BB10	BB10	BB10	rJoe	rJoe	Frank	Frank
n							
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	BB10	BVN	BB10	BVN	rJoe	rJoe	rGumbel
n							
100	5	14	28	10	29	31	10
300	27	29	43	22	49	40	16
500	39	39	60	31	55	63	31

Note: rCopula: reflected copula; 1rCopula: 1-reflected copula; 2rCopula: 2-reflected copula.

8. Discussion

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 nonlinear (e.g., Rizopoulos & Moustaki, 2008), semi-parametric (e.g., Gruhl, Erosheva, & Crane, 2013), or nonparametric models (e.g., Kelava, Kohler, Krzyżak, & Schaffland, 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, called the factor copula model (Murray, Dunson, Carin, & Lucas, 2013), is restricted to the MVN copula like the model proposed by Gruhl et al. (2013), hence it has an additive latent structure.

We have shown that the 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 nonlinear dependence on the latent variables. The factor copula can provide flexible reflection asymmetric tail and nonlinear dependence as it is a truncated canonical vine copula (Brechmann et al., 2012) rooted at the latent variables. Joe, Li, and Nikoloulopoulos (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 at level 1 to have (tail) dependence and it is not necessary for the conditional bivariate copulas at levels $2,\ldots,d-1$ to have tail

dependence. The one-factor copula has bivariate copulas with tail dependence at the first level and independence copulas at all the remaining levels of the vine (truncated after the first level). The two-factor copula has bivariate copulas with tail dependence at the first and second level and independence copulas at all the remaining levels (truncated after the second level). Hence, the tail dependence among the latent variables and each of the observed variables is inherited by the tail dependence among the observed variables.

Even in cases where the effect of misspecifying the bivariate linking copula choice 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 2.1, and is a property to consider when choosing among different families of copulas and hence affects prediction. Rabe-Hesketh, Pickles, and Skrondal (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. The 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 inherited by the dependence between the factor scores and each of the observed variables. Hence, factor scores are fairly different than those for the standard factor model if the sample size is sufficient. Figure 3 demonstrates these differences by revisiting the political-economic

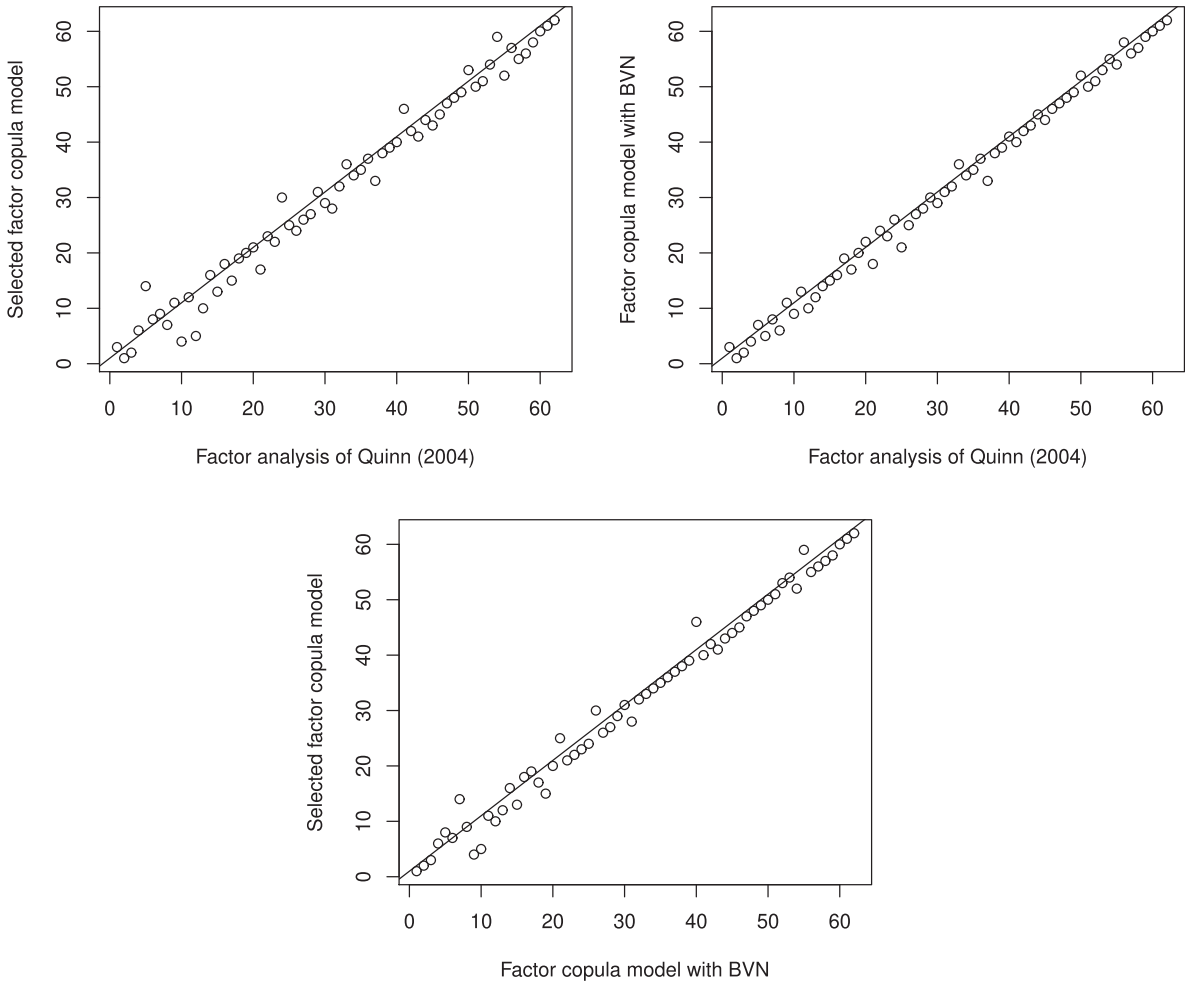

Figure 3. Comparison of the political-economic risk rankings obtained via our selected model, the standard factor model, and the mixed-data factor analysis of Quinn (2004).

data set in Section 6.1 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, that is, allows the margins to be quite free and not restricted by the normal distribution. The differences in the lower panel graph are solely due the misspecification of the latent variable distribution.

As stated by many researchers (e.g., Rabe-Hesketh & Skrondal, 2001, 2004), the major difficulty for 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 the 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 two-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 two-factor copula model is near-identifiable with $2d$ bivariate linking copulas, as it as been demonstrated by Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015).

Acknowledgements

We would like to thank the referees and Professor Harry Joe (University of British Columbia) for their careful reading and comments that led to an improved presentation, and Dr Irina Irincheeva (University of Bern) and Professor Marc Genton (King Abdullah University of Science and Technology) for sharing the Swiss Consumption Survey data set. The simulations presented in this paper were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Sayed H. Kadhem (Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing - review & editing) Aristidis K. Nikoloulopoulos (Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Validation; Visualization; Writing - original draft; Writing – review & editing).

Data availability statement

Our modelling framework is implemented in the package *FactorCopula* (Kadhem & Nikoloulopoulos, 2020) within the open source statistical environment R (R Core Team,

2020). All the analyses presented in Sections 6.1 and 6.2 are given as code examples in the package.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, Wiley Series in Probability and Statistics, 2nd edition, Hoboken, NJ: John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470594001

Brechmann, E. C., Czado, C., & Aas, K. (2012). Truncated regular vines in high dimensions with applications to financial data. *Canadian Journal of Statistics*, *40*(1), 68–85. https://doi.org/10.1002/cjs.10141

Genest, C., Ghoudi, K., & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*, 543–552. https://doi.org/10.1093/biomet/82.3.543

Genest, C., & Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, *37*, 475–515. https://doi.org/10.2143/AST.37.2.2024077

Gruhl, J., Erosheva, E. A., & Crane, P. K. (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *Annals of Applied Statistics*, *7*, 2361–2383. https://doi.org/10.1214/13-AOAS675

He, J., Li, H., Edmondson, A. C., Rader, D. J., & Li, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics*, *13*, 497–508. https://doi.org/10.1093/biostatistics/kxr025

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, *1*(1), 265–283. https://doi.org/10.1214/07-AOAS107

Hua, L., & Joe, H. (2011). Tail order and intermediate tail dependence of multivariate copulas. *Journal of Multivariate Analysis*, *102*, 1454–1471. https://doi.org/10.1016/j.jmva.2011.05.011

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B*, *66*, 893–908. https://doi.org/10.1111/j.1467-9868.2004.05627.x

Irincheeva, I., Cantoni, E., & Genton, M. (2012a). A non-Gaussian spatial generalized linear latent variable model. *Journal of Agricultural, Biological, and Environmental Statistics*, *17*, 332–353. https://doi.org/10.1007/s13253-012-0099-5

Irincheeva, I., Cantoni, E., & Genton, M. G. (2012b). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, *39*, 663–680. https://doi.org/10.1111/j.1467-9469.2011.00777.x

Jiryaie, F., Withanage, N., Wu, B., & de Leon, A. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation*, *86*, 1643–1659. https://doi.org/10.1080/00949655.2015.1077386

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, *46*(2), 262–282. https://doi.org/10.1006/jmva.1993.1061

Joe, H. (1997). *Multivariate models and dependence concepts*. London, UK: Chapman & Hall.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*, 401–419. https://doi.org/10.1016/j.jmva.2004.06.003

Joe, H. (2011). Tail dependence in vine copulae. In D. Kurowicka & H. Joe (Eds.), *Dependence modeling: Vine copula handbook* (pp. 165–187). Singapore: World Scientific.

Joe, H. (2014). *Dependence modelling with copulas*. Boca Raton, FL: Chapman and Hall/CRC.

Joe, H., Li, H., & Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, *101*(1), 252–270. https://doi.org/10.1016/j.jmva.2009.08.002

Kadhem, S. H., & Nikoloulopoulos, A. K. (2020). *FactorCopula: Factor copula models for mixed continuous and discrete data*. R package version 0.5. Retrieved from http://CRAN.R-project.org/package=FactorCopula

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23* (3), 187–200. https://doi.org/10.1007/BF02289233

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, *56*, 4243–4258. https://doi.org/10.1016/j.jmva.2016.10.006

Kelava, A., Kohler, M., Krzyżak, A., & Schaffland, T. F. (2017). Nonparametric estimation of a latent variable model. *Journal of Multivariate Analysis*, *154*, 112–134. https://doi.org/10.1016/j.jmva.2016.10.006

Kim, G., Silvapulle, M. J., & Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, *51*, 2836–2850. https://doi.org/10.1016/j.csda.2006.10.009

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

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, *15*(3), 209–225. https://doi.org/10.2307/3314912

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*(1), 89–105. https://doi.org/10.1007/BF02294660

Ma, Y., & Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B*, *72*, 475–495. https://doi.org/10.1111/j.1467-9868.2010.00741.x

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

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305–328. https://doi.org/10.1080/00273171.2014.911075

Montanari, A., & Viroli, C. (2010). A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, *37*, 473–487. https://doi.org/10.1080/02664760902736737

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. https://doi.org/10.1111/j.2044-8317.1996.tb01091.x

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391–411. https://doi.org/10.1007/BF02296153

Moustaki, I., & Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, *101*, 644–653. https://doi.org/10.1198/016214505000001320

Murray, J. S., Dunson, D. B., Carin, L., & Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, *108*, 656–665. https://doi.org/10.1080/01621459.2012.762328

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Nash, J. (1990). *Compact numerical methods for computers: Linear algebra and function minimisation* (2nd edition). Bristol and New York: Adam Hilger.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, *143*, 1923–1937. https://doi.org/10.1016/j.jspi.2013.06.015

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. https://doi.org/10.1007/s00477-015-1060-2

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*, 2270–2286. https://doi.org/10.1177/0962280215596769

Nikoloulopoulos, A. K., & Joe, H. (2015). Factor copula models for item response data. *Psychometrika*, *80*(1), 126–150. https://doi.org/10.1007/s11336-013-9387-4

Nikoloulopoulos, A. K., Joe, H., & Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics and Data Analysis*, *56*, 3659–3673. https://doi.org/10.1016/j.csda.2010.07.016

Nikoloulopoulos, A. K., & Karlis, D. (2008). Copula model evaluation based on parametric bootstrap. *Computational Statistics and Data Analysis*, *52*, 3342–3353. https://doi.org/10.1016/j.csda.2007.10.028

North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History*, *49*, 803–832. https://doi.org/10.1017/S0022050700009451

Panagiotelis, A., Czado, C., Joe, H., & Stöber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics and Data Analysis*, *106*, 138–152. https://doi.org/10.1016/j.csda.2016.09.007

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, *12*, 338–353. https://doi.org/10.1093/pan/mph022

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, *3*, 215–232. https://doi.org/10.1191/1471082X03st056oa

Rabe-Hesketh, S., & Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, *57*, 1256–1263. https://doi.org/10.1111/j.0006-341X.2001.1256_1.x

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

Rizopoulos, D., & Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 415–438. https://doi.org/10.1348/000711007X213963

Shen, C., & Weissfeld, L. (2006). A copula model for repeated measurements with non-ignorable non-monotone missing outcome. *Statistics in Medicine*, *25*, 2427–2440. https://doi.org/10.1002/sim.2355

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, *51*, 1384–1399. https://doi.org/10.2307/2533269

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.

Smith, M. S., & Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, *107*, 290–303. https://doi.org/10.1080/01621459.2011.644501

Song, P.-X.-K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, *65*(1), 60–68. https://doi.org/10.1111/j.1541-0420.2008.01058.x

Stöber, J., Hong, H. G., Czado, C., & Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics and Data Analysis*, *88*, 28–39. https://doi.org/10.1016/j.csda.2015.02.001

Stroud, A., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Takane, Y., de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. https://doi.org/10.1007/BF02294363

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Wedel, M., & Kamakura, W. A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, *66*, 515–530. https://doi.org/10.1007/BF02296193

Zilko, A. A., & Kurowicka, D. (2016). Copula in a multivariate mixed discrete–continuous model. *Computational Statistics and Data Analysis*, *103*, 28–55. https://doi.org/10.1016/j.csda.2016.02.017

Received 20 December 2019; revised version received 16 November 2020