Factor copula models for mixed data

Sayed H. Kadhem* Aristidis K. Nikoloulopoulos †

Abstract

We develop factor copula models for analysing 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 non-linear dependence. They can be explained as conditional independence models with latent variables that don’t necessarily have an additive latent structure. We focus on important issues that would interest 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 re-analysing three mixed response datasets. Our study suggests that there can be a substantial improvement over the standard factor model for mixed data and makes the argument for moving to factor copula models.

Key Words: Conditional independence; Goodness-of-fit; Latent variable models; Model selection; Tail dependence/asymmetry; Canonical vines.

1 Introduction

It is very common in social science, e.g., in surveys, to deal with datasets 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 as a manifestation of underlying continuous variables that follow the normal distribution (e.g., Muthén 1984; Lee et al. 1992; Quinn 2004), and, the response function approach that postulates distributions on the observed variables conditional on the latent variables usually from the exponential family (e.g., Moustaki 1996; Moustaki and Knott 2000; Wedel and Kamakura 2001; Huber et al. 2004; Moustaki and Victoria-Feser 2006). The former method assumes that the continuous and underlying variables follow a multivariate normal (MVN) distribution, while the latter assumes that the observed variables are conditionally independent given the latent variables which again

*School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, U.K.
†Correspondence to: a.nikoloulopoulos@uea.ac.uk, Aristidis K. Nikoloulopoulos, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.
follow a MVN distribution. The underlying variable approach calls the MVN distribution as a latent model for the discrete responses, hence requires multidimensional integrations (Nikoloulopoulos, 2013, 2016). The response function approach achieves dimension reduction with dependence coming from \( p \) latent (unobservable) variables/factors where \( p << d \), where \( d \) is the number of observed variables and hence estimation is more feasible when many categorical or discrete variables are involved. Nevertheless, both approaches are restricted to MVN assumption that is not valid if tail asymmetry or tail dependence exists in the mixed data which is a realistic scenario. Ma and Genton (2010), Montanari and Vircoli (2010), and Irincheeva et al. (2012a) stress that the MVN assumption might not be adequate, and acknowledge that the effect of misspecifying the distribution of the latent variables could lead to biased model estimates and poor fit. To this end, Irincheeva et al. (2012b) proposed a more flexible response function approach by strategically multiplying the MVN density of the latent variables by a polynomial function to achieve departures from normality.

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

Smith and Khaled (2012), Stöber et al. (2015) and Zilko and Kurowicka (2016) called vine copulas to model mixed data. Vine copulas have two major advantages over the MVN copula as emphasized in Panagiotelis et al. (2017). The first is that the computational complexity of computing the joint probability distribution function grows 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. Note in passing that they have as special case the MVN copula, if all the bivariate 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 copula models are vine copulas that involve both observed and latent variables, hence they allow flexible (tail) dependence. These types of models are more interpretable and fit better than vine copula models, when dependence can be explained through latent variables. That is theoretical concepts that cannot be measured directly such as intelligence in psychology, or welfare and poverty in economics. Furthermore, the factor copula models are closed under margins (same model with one more or less variable) whereas vine copulas without latent variables are not closed under margins. Finally, they can be explained as conditional independence models, i.e., they are a response function
approach with dependence coming from latent (unobservable) variables/factors. Our model is the most general conditional independence model with univariate parameters separated from dependence parameters and latent variables that don’t necessarily have an additive latent structure. Another mixed-variable model in the literature that is called factor copula model (Murray et al., 2013) is restricted to the MVN copula, hence has an additive latent structure and is not as general.

The remainder of the paper proceeds as follows. Section 2 introduces the factor copula models for mixed data and discusses its relationship with existing models. Estimation techniques and computational details are provided in Section 3. 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, model selection and goodness-of-fit techniques. We conclude with some discussion in Section 8.

2 The factor copula model for mixed responses

The \( p \)-factor model assumes that the mixed continuous and discrete responses \( Y = (Y_1, \ldots, Y_d) \) are conditionally independent given \( p \) latent variables \( X_1, \ldots, X_p \). In line with Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015), we will use a general copula construction, based on a set of bivariate copulas that link observed to latent variables, to specify the factor copula models for mixed continuous and discrete variables. The idea in the derivation of this \( p \)-factor model will be shown below for the 1-factor and 2-factor case. It can be extended to \( p \geq 3 \) factors or latent variables in a similar manner.

For the 1-factor model, let \( X_1 \) be a latent variable, which we assume to be standard uniform (without loss of generality). From Sklar (1959), there is a bivariate copula \( C_{X_1,j} \) such that

\[
\Pr(X_1 \leq x, Y_j \leq y) = C_{X_1,j}(x, F_j(y)) \quad \text{for } 0 \leq x \leq 1
\]

where \( F_j \) is the cumulative distribution function (cdf) of \( Y_j \). Then it follows that

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

Letting \( C_{j|X_1}(F_j(y)|x) = \partial C_{X_1,j}(x, F_j(y))/\partial x \) for shorthand notation and \( y = (y_1, \ldots, y_d) \) be realizations of \( Y \), the density\(^\dagger\) for the 1-factor model is

\[
f_Y(y) = \int_0^1 \prod_{j=1}^d f_{j|X_1}(y_j|x) \, dx,
\]

\(^\dagger\)We mean the density of \( Y \) w.r.t. the product measure on the respective supports of the marginal variables. For discrete margins with integer values this is the counting measure on the set of possible outcomes, for continuous margins we consider the Lebesgue measure in \( \mathbb{R} \).
where
\[ f_{j|X_1}(y|x) = \begin{cases} C_{j|X_1}(F_j(y + 1)|x) - C_{j|X_1}(F_j(y)|x) & \text{if } Y_j \text{ is discrete;} \\ c_{X_1,j}(x, F_j(y))f_j(y) & \text{if } Y_j \text{ is continuous,} \end{cases} \]
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 2-factor model, consider two latent variables \( X_1, X_2 \) that are, without loss of generality, independent uniform \( U(0,1) \) random variables. Let \( C_{X_1,j} \) be defined as in the 1-factor model, and let \( C_{X_2,j} \) be a bivariate copula such that
\[
\Pr(X_2 \leq x_2, Y_j \leq y|X_1 = x_1) = C_{X_2,j}(x_2, F_j|X_1(y|x_1)),
\]
where \( F_j|X_1 \) is given in (1). Then for \( 0 \leq x_1, x_2 \leq 1 \),
\[
\Pr(Y_j \leq y|X_1 = x_1, X_2 = x_2) = \frac{\partial}{\partial x_2} \Pr(X_2 \leq x_2, Y_j \leq y|X_1 = x_1)
\]
\[
= \frac{\partial}{\partial x_2} C_{X_2,j}(x_2, F_j|X_1(y|x_1)) = C_{j|X_2}(F_j|X_1(y|x_1)|x_2).
\]
The density for the 2-factor model is
\[
f_Y(y) = \int_0^1 \int_0^1 \prod_{j=1}^d f_{X_2,j|X_1}(x_2, y_j|x_1) \, dx_1 dx_2,
\]
where
\[
f_{X_2,j|X_1}(x_2, y|x_1) = \begin{cases} C_{j|X_2}(F_j|X_1(y|x_1)|x_2) - C_{j|X_2}(F_j|X_1(y - 1|x_1)|x_2) & \text{if } Y_j \text{ is discrete;} \\ c_{X_2,j}(F_j|X_1(y|x_1), x_2) c_{X_1,j}(x_1, F_j(y)) f_j(y) & \text{if } Y_j \text{ is continuous.} \end{cases}
\]

Note that the copula \( C_{X_1,j} \) links the \( j \)th response to the first latent variable \( X_1 \), and the copula \( C_{X_2,j} \) links the \( j \)th 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

In this subsection, we provide choices of parametric bivariate copulas that can be used to link the latent with observed variables. We will consider copula families that have different tail dependence (Joe, 1993) or tail order (Hua and 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 \( 0 \leq u_1, u_2 \leq 1 \). Otherwise, it is reflection asymmetric often with more probability in the joint upper tail or joint lower tail. Upper tail dependence means that \( c(1-u, 1-u) = O(u^{-1}) \) as \( u \to 0 \) and lower tail dependence means that \( c(u, u) = O(u^{-1}) \) as \( u \to 0 \). If \( (U_1, U_2) \sim C \) for a bivariate copula \( C \), then \( (1-U_1, 1-U_2) \sim \tilde{C} \), where \( \tilde{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1-u_1, 1-u_2) \) 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 exists $\kappa_L(C) > 0$ and some $L(u)$ that is slowly varying at $0^+$ (i.e., $\frac{L(u)}{L(u)} \sim 1$, 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 $(U_1, U_2)$, i.e., $C(1-u, 1-u) \sim u^{\kappa_U(C)}L^*(u)$ as $u \to 0^+$, where $C$ is the survival function of the copula and $L^*(u)$ is a slowly varying function. With $\kappa = \kappa_L$ or $\kappa_U$, 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 asymptotically a constant.

After briefly providing 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 $\theta$ 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 $\theta$ is the copula parameter;
  - the Joe copula with $\kappa_L = 2$ and $\kappa_U = 1$.
- Reflection symmetric copulas with tail dependence, such as the $t_\nu$ copula with $\kappa_L = \kappa_U = 2 - 2T_{\nu+1}(1/(1 - \theta)/(1 + \theta))$, where $\theta$ is the correlation parameter of the bivariate $t$ distribution with $\nu$ degrees of freedom, and $T_\nu$ is the univariate $t$ cdf with $\nu$ degrees of freedom.
- 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 the BB8 and BB10 copulas.

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

Figure 1: Contour plots of bivariate copulas with standard normal margins and dependence parameters corresponding to Kendall’s \(\tau\) value of 0.5.

| BVN          | Frank     | \(t_3\)        |
|--------------|-----------|-----------------|
| ![Contour Plot](BVN.png) | ![Contour Plot](Frank.png) | ![Contour Plot](t3.png) |

| Joe          | reflected Joe | 1-reflected Joe |
|--------------|---------------|-----------------|
| ![Contour Plot](Joe.png) | ![Contour Plot](reflected Joe.png) | ![Contour Plot](1-reflected Joe.png) |

| BB10         | reflected BB10 | 2-reflected BB10 |
|--------------|-----------------|------------------|
| ![Contour Plot](BB10.png) | ![Contour Plot](reflected BB10.png) | ![Contour Plot](2-reflected BB10.png) |

In Figure 1, to depict the concepts of reflection symmetric or asymmetric tail dependence or quadrant tail
independence, we plot contour plots of the corresponding copula densities with standard normal margins and
dependence parameters corresponding to Kendall’s $\tau$ value of 0.5.

3 Estimation

Suppose that data are $y_{ij}$, $j = 1, \ldots, d$, $i = 1, \ldots, n$, where $i$ is an index for individuals or clusters and $j$ is
an index for the within-cluster measurements. For continuous random variables, we estimate each marginal
distribution non-parametrically by the empirical distribution function of $Y_j$, viz.

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

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

Nevertheless, rank-based methods cannot be used for discrete variables with copulas (Genest and Nešlehová,
2007). Hence,

- For an ordinal response variable $Y_j$ we use a univariate probit model (Agresti, 2010, Section 3.3.2). Let $Z_j$ be a standard normal latent variable, such that

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

where $K_j$ is the number of categories of $Y_j$ (without loss of generality, we assume $\alpha_{0j} = -\infty$ and $\alpha_{K_j} = \infty$). From this definition, the ordinal response $Y_j$ is assumed to have density

$$f_j(y_j; \gamma_j) = \Phi(\alpha_{y_{j,j}}) - \Phi(\alpha_{y_{j-1,j}}),$$

where $\gamma_j = \{a_{1j}, \ldots, a_{K_j-1,j}\}$ is the vector with the univariate cutpoints.

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

$$f_j(y_j; \gamma_j) = \frac{\Gamma(\xi^{-1} + y_j)}{\Gamma(\xi^{-1}) y_j!} \frac{\mu^\xi y_j^\xi}{(1 + \xi^{-1})^{\xi^{-1} + y_j}}, \quad y_j = 0, 1, 2, \ldots, \mu > 0, \xi > 0,$$

where $\gamma_j = \{\mu_j, \xi_j\}$ is the vector with the mean and dispersion parameters. In the limit $\xi \to 0$ the
negative binomial reduces to Poisson, which belongs to the exponential family of distributions and it is the only distribution for count data that existing latent variable models for mixed data can accommodate.
To this end, for a discrete random variable $Y_j$, we approach estimation by maximizing the univariate log-likelihoods

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

over the vector of the univariate parameters $\gamma_j$. That is equivalent with the first step of the IFM method in Joe (1997, 2005).

After estimating the univariate marginal distributions we proceed to estimation of the dependence parameters. For the 1-factor and 2-factor models, we let $C_{X_{1,j}}$ and $C_{X_{2,j}}$ be parametric bivariate copulas, say with dependence parameters $\theta_j$ and $\delta_j$, respectively. Let also $\theta = \{\gamma_j, \theta_j : j = 1, \ldots, d\}$ and $\theta = \{\gamma_j, \theta_j, \delta_j : j = 1, \ldots, d\}$ to denote the set of all parameters for the 1- and 2-factor model, respectively. Estimation can be achieved by maximizing the joint log-likelihood

$$\ell_Y(\theta) = \sum_{i=1}^{n} \log f_Y(y_{i1}, \ldots, y_{id}; \theta),$$

over the copula parameters $\theta_j$ or $\delta_j$, $j = 1, \ldots, d$ with the univariate parameters/distributions fixed as estimated at the first step of the proposed two-step estimation approach. The estimated parameters can be obtained by using a quasi-Newton Nash (1990) method applied to the logarithm of the joint likelihood. This numerical method requires only the objective function, i.e., the logarithm of the joint likelihood, while the gradients are computed numerically and the Hessian matrix of the second order derivatives is updated in each iteration. The standard errors (SE) of the estimates can be obtained via the gradients and the Hessian computed numerically during the maximization process. These SEs are adequate to assess the flatness of the log-likelihood. Proper SEs that account for the estimation of univariate parameters can be obtained by maximizing the joint likelihood in (4) at one step over $\theta$.

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

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

where \{x_q : q = 1, \ldots, n_q\} are the quadrature points and \{w_q : q = 1, \ldots, n_q\} are the quadrature weights. To compute two-dimensional integrals for the 2-factor model, the approximation uses Gauss-Legendre quadrature...
points in a double sum:

\[ f_Y(y) = \int_0^1 \int_0^1 \prod_{j=1}^d f_{X_{2j}|X_1}(x_{2j}, y_j | x_1) \, dx_1 \, dx_2 \]

\[ \approx \sum_{q_1=1}^{n_q} \sum_{q_2=1}^{n_q} w_{q_1} w_{q_2} \prod_{j=1}^d f_{X_{2j}|X_1}(x_{q_2}, y_j | x_{q_1}). \]

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 an heuristic method that automatically selects the bivariate parametric copula families that link the observed to the latent variables. For multivariate mixed data, it is infeasible to estimate all possible combinations of bivariate parametric copula families, and compare them on the basis of information criteria. We develop an algorithm that can quickly select a factor copula model that accurately captures the (tail) dependence features in the data on hand. The linking copulas at each factor are selected with a sequential algorithm under the initial assumption that linking copulas are Frank, and in then sequentially copulas with non-tail quadrant independence are assigned to any of pairs where necessary to account for tail asymmetry (discrete data) or tail dependence (continuous data). Tail dependence of discrete variables is not clear, because maxima of discrete variables may not converge to a nondegenerate distribution (Feidt et al., 2010).

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

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

2. Fit the 1-factor copula model with Frank copulas to link each of the \( d \) observed variables with the latent variable, i.e., maximise the log-likelihood function of the factor copula model in (4) over the vector of copula parameters \( (\theta_1, \ldots, \theta_d) \).

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

4. For \( j = 1, \ldots, d \),
(a) Fit all the possible 1-factor copula models iterating over all the copula candidates for the \( j \)th variable.

(b) Select the copula family that corresponds to the lowest information criterion, say the Akaike, that is \( \text{AIC} = -2 \times \ell + 2 \times \# \text{copula parameters} \).

(c) Fix the selected linking copula family for the \( j \)th variable.

(d) Iterate through step (a) – (c) to select the copulas that link all the observed variables to the 1st factor.

For more than one factors 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 the Vuong’s test (Vuong, 1989) to show if a factor copula model provides better fit than the standard factor model with a latent additive structure, that is a factor copula model with BVN bivariate linking copulas (Krupskii and Joe, 2013; Nikoloulopoulos and Joe, 2015). The Vuong’s test is the sample version of the difference in Kullback-Leibler divergence between two models and can be used to differentiate two parametric models which could be non-nested. This test has been used extensively in the copula literature to compare vine copula models (e.g., Brechmann et al. 2012; Joe 2014; Nikoloulopoulos 2017). We provide specific details in Section 5.1.

Furthermore, to assess the overall goodness-of-fit of the factor copula models for mixed data, we will use appropriately the limited information \( M_2 \) statistic (Maydeu-Olivares and Joe, 2006). The \( M_2 \) statistic has been developed for goodness-of-fit testing in multidimensional contingency tables. Nikoloulopoulos and Joe (2015) has used the \( M_2 \) statistic to assess the goodness-of-fit of factor copula models for ordinal data. We build on the aforementioned papers and propose a methodology to assess the overall goodness-of-fit of factor copula models for mixed continuous and discrete responses. Since the \( M_2 \) statistic has been developed for multivariate ordinal data, we propose to first transform the continuous and count variables to ordinal and then calculate the \( M_2 \) statistic at the maximum likelihood estimate before discretization. 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 the Vuong’s test for comparing parametric models (Vuong, 1989). Assume that we have Models 1 and 2 with parametric densities \( f_Y^{(1)} \) and \( f_Y^{(2)} \), respectively. We can compare
\[
\Delta_{1f_Y} = n^{-1} \left[ \sum_{i=1}^{n} \left\{ E_{f_Y} \log f_Y(y_i) - E_{f_Y} \log f_Y^{(1)}(y_i; \theta_1) \right\} \right], \\
\Delta_{2f_Y} = n^{-1} \left[ \sum_{i=1}^{n} \left\{ E_{f_Y} \log f_Y(y_i) - E_{f_Y} \log f_Y^{(2)}(y_i; \theta_2) \right\} \right].
\]

where \( \theta_1, \theta_2 \) are the parameters in Models 1 and 2, respectively, that lead to the closest Kullback-Leibler divergence to the true \( f_Y \); equivalently, they are the limits in probability of the MLEs based on Models 1 and 2, respectively.

Model 1 is closer to the true \( f_Y \), i.e., is the better fitting model if \( \Delta = \Delta_{1f_Y} - \Delta_{2f_Y} < 0 \), and Model 2 is the better fitting model if \( \Delta > 0 \). The sample version of \( \Delta \) with MLEs \( \hat{\theta}_1, \hat{\theta}_2 \) is

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

where \( D_i = \log \left[ \frac{f_Y^{(2)}(y_i; \hat{\theta}_2)}{f_Y^{(1)}(y_i; \hat{\theta}_1)} \right] \). Vuong (1989) has shown that asymptotically that

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

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

5.2 \( M_2 \) goodness-of-fit statistic

In this subsection to self maintain the paper we provide the form of the limited information \( M_2 \) statistic in Maydeu-Olivares and Joe (2006). Assume \( d \) ordinal variables \( Y_1, \ldots, Y_d \) where the \( j \)th \( (1 \leq j \leq d) \) variable consists of \( K_j \geq 2 \) categories labelled as \( 0, 1, \ldots, K_j - 1 \). Consider the set of univariate and bivariate residuals that do not include category 0. This is a residual vector of dimension

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

For a factor copula model with parameter vector \( \theta \) of dimension \( q \), let \( \pi_2(\theta) = (\hat{\pi}_1(\theta)^\top, \hat{\pi}_2(\theta)^\top)^\top \) be the column vector of the model-based marginal probabilities with \( \hat{\pi}_1(\theta) \) the vector of univariate marginal probabilities, and \( \hat{\pi}_2(\theta) \) the vector of bivariate marginal probabilities. Also, let \( \mathbf{p}_2 = (\hat{\mathbf{p}}_1^\top, \hat{\mathbf{p}}_2^\top)^\top \) be the vector of the observed sample proportions, with \( \hat{\mathbf{p}}_1 \) the vector of univariate marginal proportions, and \( \hat{\mathbf{p}}_2 \) the vector of the bivariate marginal proportions.

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

\[
M_2 = M_2(\hat{\theta}) = n(\mathbf{p}_2 - \pi_2(\hat{\theta}))^\top C_2(\hat{\theta}) (\mathbf{p}_2 - \pi_2(\hat{\theta})),
\]

where
with
\[
C_2(\theta) = \Xi_2^{-1} - \Xi_2^{-1} \Delta_2 (\Delta_2^\top \Xi_2^{-1} \Delta_2)^{-1} \Delta_2^\top \Xi_2^{-1} = \Delta_2^{(c)} (\Delta_2^{(c)} \top \Xi_2 \Delta_2^{(c)})^{-1} \Delta_2^{(c)} \top,
\]
(6)

where \(\Delta_2 = \partial \pi_2(\theta)/\partial \theta^\top\) is an \(s \times q\) matrix with the derivatives of all the univariate and bivariate marginal probabilities with respect to the model parameters, \(\Delta_2^{(c)}\) is an \(s \times (s-q)\) orthogonal complement to \(\Delta_2\), such that \((\Delta_2^{(c)} \top \Delta_2 = 0\), and \(\Xi_2 = \text{diag}(\pi_2(\theta)) - \pi_2(\theta) \pi_2(\theta)^\top\) is the \(s \times s\) covariance matrix of all the univariate and bivariate marginal sample proportions, excluding category 0. Due to equality in (6), \(C_2\) is invariant to the choice of orthogonal complement. The limited information statistic \(M_2\) has a null asymptotic distribution that is \(\chi^2\) with \(s-q\) degrees of freedom when the estimate \(\hat{\theta}\) is \(\sqrt{n}\)-consistent. For details on the computation of \(\Xi_2\) and \(\Delta_2\) 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 re-analysing three mixed response datasets. We 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 also fit the factor copula models with the comprehensive bivariate parametric copula families that allow for reflection symmetric dependence; these are the BVN, Frank, and \(t_\nu\) copulas. For \(t_\nu\) copulas, we summarize the choice of integer \(\nu\) with the largest log-likelihood. To make it easier to compare strengths of dependence, we convert the estimated parameters to Kendall’s \(\tau\)’s in \((-1, 1)\) via the relations in Joe (2014, Chapter 4); SEs are also converted via the delta method.

If the number of parameters is not the same between the models, we use the AIC as a rough diagnostic measure for goodness-of-fit between the models, otherwise we use the likelihood at the maximum likelihood estimates.

We further compute the Vuong’s tests with Model 1 being the factor copula model with BVN copulas, that is the standard factor model, to reveal if any other factor copula model provides better fit than the standard model. For the standard 2-factor model, to obtain a unique solution we must impose sufficient constraints. One parameter for the second factor can be set to zero and the likelihood can be maximized with respect to other \(2d-1\) parameters. We report the varimax transform (Kaiser, 1958) of the loadings (a reparametrization of \(2d\) parameters), converted to factor copula parameters via the relations
\[
\theta_j = \beta_{j1}, \quad \delta_j = \frac{\beta_{j2}}{(1 - \beta_{j1}^2)^{1/2}},
\]
(7)

where \(\beta_{j1}\) and \(\beta_{j2}\) are the loadings at the first and second factor, respectively. 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 $\Delta_2$ having one less column.

6.1 Political–economic dataset

Quinn (2004) considered measuring the (latent) political–economic risk of 62 countries, for the year 1987, using 5 mixed variables, namely the continuous variable black-market premium in each country (used as a proxy for illegal economic activity), the continuous variable productivity as measured by real gross domestic product per worker in 1985 international prices, the binary variable independence of the national judiciary (1 if the judiciary is judged to be independent and 0 otherwise), and the ordinal variables measuring the lack of expropriation risk and lack of corruption. The dataset and its complete description can be found in Quinn (2004) or in the R package MCMCpack (Martin et al., 2011).

Table 1 gives the estimated parameters, their standard errors (SE) in Kendalls $\tau$ scale, joint log-likelihoods, the 95% CIs of Vuong’s tests, and the $M_2$ statistics for the 1-factor copula models. Table 1 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 $t_\nu$ copulas.

*Table 1: Estimated parameters, their standard errors (SE) in Kendall’s $\tau$ scale, joint log-likelihoods, the 95% CIs of Vuong’s statistics, and the $M_2$ statistics for the one-factor copula models for the political-economic risk data.*

| 1-factor | 1-factor BVN¶ | 1-factor $t_5$ | Frank | Selected copulas | 1-factor $t_\nu$ |
|----------|----------------|----------------|-------|-----------------|-----------------|
|          | $\hat{\tau}$ | SE             | $\hat{\tau}$ | SE             | $\hat{\tau}$ | SE             |
| BM       | -0.50          | 0.06           | -0.51  | 0.07            | -0.49          | 0.06           |
| GDP      | 0.57           | 0.05           | 0.57   | 0.06            | 0.58           | 0.06           |
| IJ       | 0.80           | 0.09           | 0.81   | 0.09            | 0.75           | 0.09           |
| XPR      | 0.66           | 0.06           | 0.68   | 0.07            | 0.66           | 0.06           |
| CRP      | 0.71           | 0.06           | 0.70   | 0.06            | 0.72           | 0.06           |

|          | $\ell$         | Vuong 95% CI   | $M_2$  | df              | p-value |
|----------|----------------|----------------|-------|-----------------|---------|
|          | -165.15        | (-0.05,-0.01)  | 179.2 | 134             | 0.01    |
|          | -166.25        | (-0.08,0.08)   | 187.4 | 134             | < 0.001 |
|          | -164.89        | (0.07,0.35)    | 177.6 | 134             | 0.001   |
|          | -151.98        | (0.07,0.35)    | 129.2 | 134             | 0.60    |

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

In all the fitted models the estimated Kendall’s $\tau$’s are similar. Kendall’s $\tau$ only accounts for the dependence dominated by the middle of the data, and it is expected to be similar amongst different families of copulas.
However, the tail dependence and tail order vary, as explained in Section 2.1, and they are properties to consider when choosing amongst different families of copulas (Nikoloulopoulos and Karlis, 2008).

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

### 6.2 General social survey

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

Table 2 gives the estimated parameters, their standard errors (SE) in Kendalls $\tau$ scale, the joint log-likelihoods, the 95% CIs of Vuong’s tests, and the $M_2$ statistics for the 1-factor and 2-factor copula models. The best fit for the 1-factor model is based on the bivariate copulas selected by the proposed algorithm, where there is improvement over the factor copula model with BVN copulas according to the 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 was categorised into 4 age groups (18–24, 25–44, 45–64, and 65+), income was grouped into five categories that conforms to approximately equal percentiles, number of children of the survey respondent and respondent’s parents were grouped into 4 and 8 categories, respectively, as there were very few respondents and respondent’s parents with more than 3 and 7 children, respectively.

The 2-factor copula models with BVN, $t_\nu$, and Frank copulas provide some improvement over the 1-factor
Table 2: Estimated parameters, their standard errors (SE) in Kendall’s τ scale, joint log-likelihoods, the 95% CIs of Vuong’s statistics, and the M2 statistics for the 1- and 2-factor copula models for the general social survey dataset.

| 1-factor | BVN‡ | t9 | Frank | Selected |
|-----------|------|----|-------|----------|
| ‡ | τ | SE | ‡ | τ | SE | ‡ | τ | SE |
| income | 0.20 | 0.04 | 0.20 | 0.04 | 0.20 | 0.04 | Joe | 0.30 | 0.04 |
| age | -0.14 | 0.04 | -0.14 | 0.04 | -0.14 | 0.04 | 2-reflected Joe | -0.14 | 0.03 |
| degree | 0.40 | 0.04 | 0.39 | 0.04 | 0.38 | 0.04 | t5 | 0.46 | 0.05 |
| pincome | 0.33 | 0.03 | 0.34 | 0.04 | 0.35 | 0.04 | t5 | 0.33 | 0.04 |
| pdgree | 0.62 | 0.05 | 0.65 | 0.05 | 0.68 | 0.06 | reflected Gumbel | 0.55 | 0.05 |
| child | -0.20 | 0.04 | -0.19 | 0.04 | -0.19 | 0.04 | 2-reflected Joe | -0.14 | 0.03 |
| pchild | -0.32 | 0.03 | -0.31 | 0.04 | -0.32 | 0.04 | 2-reflected Gumbel | -0.27 | 0.03 |
| ℓ | -3425.39 | -3420.56 | -3433.83 | -3396.76 |
| Vuong 95%CI | (-0.005,-0.025) | (-0.037,0.001) | (0.026,0.097) |
| M2 | 743.74 | 715.45 | 738.76 | 665.69 |
| df | 348 | 348 | 348 | 348 |
| p-value | < 0.001 | < 0.001 | < 0.001 | < 0.001 |

| 2-factor | BVN‡ | t7 | Frank | Selected |
|-----------|------|----|-------|----------|
| ‡ | τ | SE | ‡ | τ | SE | ‡ | τ | SE |
| 1st factor | income | 0.36 | 0.32 | 0.36 | 0.51 | 0.04 | reflected Gumbel | 0.34 | 0.03 |
| | age | 0.05 | -0.08 | 0.44 | 0.18 | 0.05 | reflected Joe | 0.49 | 0.03 |
| | degree | 0.55 | 0.52 | 0.19 | 0.54 | 0.04 | BVN | 0.18 | 0.04 |
| | pincome | 0.27 | 0.29 | 0.07 | 0.20 | 0.04 | 1-reflected Joe | -0.13 | 0.04 |
| | pdgree | 0.48 | 0.51 | 0.10 | 0.36 | 0.05 | 1-reflected Joe | -0.13 | 0.04 |
| | child | -0.13 | -0.16 | 0.42 | 0.09 | 0.05 | reflected Joe | 0.44 | 0.04 |
| | pchild | -0.28 | -0.28 | 0.08 | -0.21 | 0.04 | Gumbel | 0.11 | 0.03 |
| 2nd factor | income | 0.38 | 0.43 | 0.26 | 0.15 | 0.06 | Gumbel | 0.40 | 0.04 |
| | age | 0.54 | 0.55 | 0.06 | 0.52 | 0.04 | 2-reflected Joe | -0.14 | 0.03 |
| | degree | 0.14 | 0.20 | 0.58 | -0.22 | 0.07 | reflected Joe | 0.65 | 0.06 |
| | pincome | -0.09 | -0.07 | 0.28 | -0.23 | 0.04 | Gumbel | 0.29 | 0.04 |
| | pdgree | -0.16 | -0.10 | 0.59 | -0.39 | 0.05 | t5 | 0.49 | 0.04 |
| | child | 0.53 | 0.52 | 0.11 | 0.53 | 0.04 | BVN | -0.24 | 0.04 |
| | pchild | 0.13 | 0.08 | 0.28 | 0.25 | 0.04 | 2-reflected Gumbel | -0.26 | 0.03 |
| ℓ | -3286.80 | -3280.31 | -3299.04 | -3235.86 |
| Vuong 95%CI | (-0.012,-0.04) | (-0.056,0.003) | (0.061,0.159) |
| M2 | 495.36 | 482.58 | 492.51 | 370.61 |
| df | 342 | 341 | 341 | 341 |
| p-value | < 0.001 | < 0.001 | < 0.001 | 0.13 |

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

copula models but according to the M2 statistic they still have a poor fit. Nevertheless, the selected 2-factor copula model using the algorithm in Section 4 shows improvement over the standard factor copula model according to the Vuong’s statistic and better fit according to the M2 statistic; it changes a p-value < 0.001 to one > 0.10.
For the 2-factor model based on the proposed algorithm for model selection, note that, without the need for a varimax rotation, the unique loading parameters ($\hat{\tau}$'s converted to normal copula parameters $\hat{\theta}_j$'s and $\hat{\delta}_j$'s and then to loadings using the relations in (7)) 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 $n = 9660$ 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 count variable (the number of bicycles in possession of the household). Irincheeva et al. (2012b), with simple descriptive statistics such as scatter plots of the original data, have shown that these mixed responses have reflection asymmetric dependence, and fitted their latent variable approach with one and two latent variables. In Figure 2 we depict the bivariate normal scores plots (Nikoloulopoulos et al., 2012) for the continuous data. The continuous data are converted to normal scores using the normal quantiles of their empirical distributions. Note that bivariate normal scores plots are better for assessing tail asymmetry and tail dependence. With a bivariate normal scores plot one can check for deviations from the elliptical shape that would be expected with the BVN copula, and hence assess if tail asymmetry and tail dependence exists on the data. 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 3 gives the estimated parameters, their standard errors (SE) in Kendalls tau scale, the joint log-likelihoods, the 95% CIs of Vuong’s tests, and the $M_2$ statistics for the 1-factor and 2-factor copula models. The best fitted 1- and 2-factor models result when we use BB10 copulas with asymmetric quadrant tail independence to link the latent variable to each of the continuous observed variables and copulas with lower tail dependence to link the latent variables to the discrete observed variables. Once again the one-factor copula model is not adequate to explain the dependence amongst the mixed responses based on the $M_2$ statistic (Table 3, 1-factor). To apply the $M_2$ statistic, we transformed the continuous to ordinal variables with 3 categories (equally weighted) and the count variable to ordinal with 6 categories (5 bicycles or more are grouped into a single category).

While it is revealed that the selected 2-factor copula model is the best model (lowest AIC) and there is sub-
stantial improvement over the standard 2-factor model, it is not apparent from the $M_2$ statistic that the response patterns are satisfactorily explained by even 2 latent variables. This is not surprising since one should expect discrepancies between the postulated parametric model and the population probabilities, when the sample size is sufficiently large (Maydeu-Olivares and Joe, 2014). In Table 4 we list the maximum deviations of observed and expected counts for each bivariate margin, that is, $D_{j_1,j_2} = n \max_{y_1,y_2} |p_{j_1,j_2,y_1,y_2} - \pi_{j_1,j_2,y_1,y_2}(\hat{\theta})|$. 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 $509/9660$ maximum occurs in the BVN factor copula model, while this drops to $133/9660$ in the selected 2-factor copula model.

For the selected 2-factor model based on the proposed algorithm, note that, without the need for a varimax
### Table 3: Estimated parameters, their standard errors (SE) in Kendall’s $\tau$ scale, joint log-likelihoods, the 95% CIs of Vuong’s statistics, and the $M_2$ statistics for the 1- and 2-factor copula models for the Swiss consumption survey dataset.

|                  | $\hat{\tau}$ | $\hat{\tau_5}$ | Frank | Selected | copulas | $\hat{\tau}$ | SE |
|------------------|--------------|----------------|-------|----------|---------|--------------|----|
| **1-factor**     |              |                |       |          |         |              |    |
| food             | 0.69         | 0.72           | 0.74  | reflected | BB10    | 0.79         | 0.00 |
| clothes          | 0.53         | 0.53           | 0.53  |          | BB10    | 0.38         | 0.00 |
| leisure          | 0.47         | 0.50           | 0.50  |          | BB10    | 0.39         | 0.00 |
| dishwasher       | 0.24         | 0.25           | 0.23  |          | reflected | Joe         | 0.28 |
| car              | 0.27         | 0.30           | 0.28  |          | reflected | Joe         | 0.23 |
| motorcycle       | 0.07         | 0.06           | 0.08  |          | reflected | Joe         | 0.13 |
| bicycles         | 0.15         | 0.15           | 0.16  |          | reflected | Joe         | 0.17 |
| **2-factor**     |              |                |       |          |         |              |    |
| food             | 0.61         | 0.34           | 0.48  | BB10     | 0.38    | 0.00         |    |
| clothes          | 0.51         | 0.32           | 0.42  |          | BB10    | 0.36         | 0.01 |
| leisure          | 0.49         | 0.35           | 0.42  |          | BB10    | 0.38         | 0.01 |
| dishwasher       | 0.14         | -0.07          | 0.08  |          | reflected | Joe         | 0.19 |
| car              | 0.12         | -0.13          | 0.07  |          | reflected | Joe         | 0.09 |
| motorcycle       | 0.01         | -0.10          | -0.08 |          | Frank   | 0.02         | 0.01 |
| bicycles         | 0.07         | -0.10          | -0.05 |          | Frank   | 0.04         | 0.01 |
| **AIC**          | 55014.32     | 54257.36       | 55105.88 |         | 48935.28 |              |    |
| Vuong 95% CI     | (0.025,0.034)| (-0.014,0.005) | (0.272,0.309) |         |          |              |    |
| $M_2$            | 2777.87      | 2737.09        | 2808.49 |         | 1611.59   |              |    |
| df               | 71           | 71             | 71    |          | 68       |              |    |
| $p$-value        | < 0.001      | < 0.001        | < 0.001 |         | < 0.001   |              |    |
| **2-factor**     |              |                |       |          |         |              |    |
| food             | 0.36         | 0.66           | 0.66  | BB10     | 0.53    | 0.01         |    |
| clothes          | 0.18         | 0.46           | 0.40  | BB10     | 0.28    | 0.01         |    |
| leisure          | 0.07         | 0.41           | 0.36  |          | BB10    | 0.30         | 0.01 |
| dishwasher       | 0.33         | 0.37           | 0.26  |          | BB10    | 0.42         | 0.01 |
| car              | 0.48         | 0.46           | 0.36  |          | reflected | Joe         | 0.35 |
| motorcycle       | 0.19         | 0.15           | 0.21  |          | reflected | Joe         | 0.17 |
| bicycles         | 0.27         | 0.27           | 0.31  |          | reflected | Gumbel      | 0.27 |
| **AIC**          | 54243.96     | 53466.18       | 53514.74 |         | 46244.56  |              |    |
| Vuong 95% CI     | (0.033,0.045)| (0.027,0.045)  | (0.387,0.419) |         |          |              |    |
| $M_2$            | 1982.17      | 1887.56        | 1945.06 |         | 452.99    |              |    |
| df               | 65           | 64             | 64    |          | 59       |              |    |
| $p$-value        | < 0.001      | < 0.001        | < 0.001 |         | < 0.001   |              |    |

*: The resulting model is the same as the standard factor model.

rotation, the unique loadings show that one factor is loaded only on the discrete variables (dishwasher, car, motorcycle, and bicycles), while both factors are loaded on the continuous variables (food, clothes, and leisure).

This reveals that the one latent variable which is only associated with the continuous variables measures the expenses, while the other which is associated with all the mixed variables measures the possession.
### Table 4: Maximum deviations $D_{j_1,j_2}$ of observed and expected counts for each bivariate margin $(j_1, j_2)$ for the 1- and 2-factor copula models for the Swiss consumption survey dataset.

| $D_{j_1,j_2}$ | 1-factor model | 2-factor model |
|--------------|----------------|----------------|
|              | 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 |

### 7 Simulations

An extensive simulation study is conducted (a) to gauge the small-sample efficiency of the proposed estimation method, (b) to examine the reliability of using the heuristic algorithm to select the correct bivariate linking copulas, and (c) to study the small-sample performance of the $M_2$ statistic after transforming the continuous and count variables to categorical.

We set the type of the variables, the univariate margins and the bivariate linking copulas, along with their univariate and dependence parameters to mimic the data analysed in Section 6. The binary variables don’t have tail asymmetries, hence parametric copulas are less distinguishable. Therefore instead of binary, we simulated from ordinal with 3 categories. We randomly generated samples of size $n = \{100, 300, 500\}$ from the selected factor copula models.

Table 5 contains the resultant biases, root mean square errors (RMSE), and standard deviations (SD), 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, SDs and RMSEs.

Table 6 contains four common nominal levels of the $M_2$ statistic under the factor copula models for mixed...
Table 5: Small sample of sizes $n = \{100, 300, 500\}$ simulations ($10^4$ replications) from the selected factor copula models in Section 6 with resultant biases, root mean square errors (RMSE) and standard deviations (SD), scaled by $n$, for the estimated parameters.

### Political-economic dataset – 1-factor model

| $\tau$ | 0.51 | 0.58 | 0.60 | 0.68 | 0.74 |
|---------|------|------|------|------|------|
| $n$     | 100  | 300  | 500  | 100  | 300  | 500  | 100  | 300  | 500  | 100  | 300  | 500  |
| $n$Bias | 0.88 | 2.30 | 3.17 | -1.36| -3.39| -4.87| 0.75 | -0.55| 0.64 | 0.21 | -0.27| 0.19 |
| $n$SD  | 4.28 | 7.60 | 9.63 | 4.19 | 7.50 | 9.08 | 5.41 | 10.91| 11.98| 4.58 | 8.43 | 9.84 |
| $n$RMSE| 4.37 | 7.95 | 10.13| 4.40 | 8.23 | 10.31| 5.47 | 10.92| 12.00| 4.59 | 8.44 | 9.84 |

### General social survey – 1-factor model

| $\tau$ | -0.14 | 0.46 | 0.33 | 0.55 | -0.14 | -0.27 |
|---------|-------|------|------|------|-------|-------|
| $n$     | 100   | 300  | 500  | 100  | 300  | 500  |
| $n$Bias | -0.11 | -0.86| -1.66| -0.06| -0.10| -0.19 |
| $n$SD  | 7.46  | 12.41| 16.01| 6.55 | 11.12| 14.00 |
| $n$RMSE| 7.46  | 12.44| 16.10| 6.55 | 11.12| 14.00 |

### Swiss consumption survey – 1-factor model

| $\tau$ | 0.69 | 0.38 | 0.39 | 0.28 | 0.23 | 0.13 | 0.17 |
|---------|------|------|------|------|------|------|------|
| $n$     | 100  | 300  | 500  | 100  | 300  | 500  | 100  | 300  | 500  | 100  | 300  | 500  |
| $n$Bias | -15.95| -0.78 | -0.04| -7.85| -1.57| -3.22| 0.08 | 0.09 | 0.26 | 0.10 | 0.06 | -0.11|
| $n$SD  | 8.81  | 9.93 | 13.16| 9.58 | 6.24 | 7.98 | 9.54 | 6.52 | 8.11 | 7.69 | 13.02| 16.80 |
| $n$RMSE| 18.23 | 9.96 | 13.16| 12.38| 6.43 | 8.60 | 12.47| 6.72 | 8.73 | 7.69 | 13.02| 16.81 |

### General social survey – 2-factor model

| $\tau$ | 0.34 | 0.38 | 0.18 | -0.13 | -0.13 | 0.44 | -0.44 |
|---------|------|------|------|-------|-------|------|-------|
| $n$     | 100  | 300  | 500  | 100  | 300  | 500  | 100  |
| $n$Bias | 1.18 | -7.19| 1.40 | 0.31 | 0.19 | 1.45 | -0.44|
| $n$SD  | 16.17 | 17.21| 19.25| 18.83| 18.63| 19.05| 17.66|
| $n$RMSE| 16.21 | 18.65| 19.30| 18.84| 18.63| 19.11| 17.67|

### Swiss consumption survey – 2-factor model

| $\tau$ | 0.34 | 0.36 | 0.38 | 0.19 | 0.09 | 0.02 | 0.04 |
|---------|------|------|------|------|------|------|------|
| $n$     | 100  | 300  | 500  | 100  | 300  | 500  | 100  |
| $n$Bias | -2.31| -1.60| -0.69| -3.01| -1.04| -0.54| 2.89 |
| $n$SD  | 7.43 | 13.67| 16.12| 27.11| 25.31| 21.27| 21.31|
| $n$RMSE| 7.78 | 13.77| 16.14| 27.27| 25.33| 21.28| 21.51|

$n = 500$
data with continuous and count data transformed to ordinal with \( K = \{3, 4, 5\} \) categories. The observed levels are close to nominal levels. Hence, it is demonstrated that the \( M_2 \) statistic remains reliable for mixed data and that the information loss under discretization is minimal.

Table 6: Small sample of sizes \( n = \{100, 300, 500\} \) distribution for \( M_2 \) (10^4 replications). Empirical rejection levels at \( \alpha = \{0.20, 0.10, 0.05, 0.01\} \), degrees of freedom \( (df) \), and mean under the factor copula models. Continuous and count variables are transformed to ordinal with \( K = \{3, 4, 5\} \) categories.

| \( n = 100 \) | \( n = 300 \) | \( n = 500 \) |
|----------------|----------------|----------------|
| \( K = 3 \) | \( K = 4 \) | \( K = 5 \) | \( K = 3 \) | \( K = 4 \) | \( K = 5 \) | \( K = 3 \) | \( K = 4 \) | \( K = 5 \) |
| **Political-economic dataset – 1-factor model** | | | | | | | | |
| df | 92 | 121 | 152 | 92 | 121 | 152 | 92 | 121 | 152 |
| mean | 89.3 | 100.2 | 148.0 | 90.6 | 118.1 | 151.6 | 100.9 | 119.1 | 151.4 |
| \( \alpha = 0.20 \) | 0.183 | 0.192 | 0.197 | 0.196 | 0.194 | 0.195 | 0.196 | 0.189 | 0.190 |
| \( \alpha = 0.10 \) | 0.122 | 0.125 | 0.134 | 0.122 | 0.121 | 0.119 | 0.115 | 0.109 | 0.109 |
| \( \alpha = 0.05 \) | 0.084 | 0.089 | 0.098 | 0.076 | 0.077 | 0.077 | 0.073 | 0.071 | 0.067 |
| \( \alpha = 0.01 \) | 0.044 | 0.046 | 0.055 | 0.036 | 0.034 | 0.037 | 0.028 | 0.030 | 0.026 |

| **General social survey – 1-factor model** | | | | | | | | |
| df | 161 | 239 | 329 | 161 | 239 | 329 | 161 | 239 | 329 |
| mean | 161.5 | 240.0 | 333.0 | 160.7 | 239.1 | 329.7 | 161.3 | 240.2 | 329.6 |
| \( \alpha = 0.20 \) | 0.213 | 0.220 | 0.240 | 0.202 | 0.216 | 0.203 | 0.211 | 0.228 | 0.212 |
| \( \alpha = 0.10 \) | 0.110 | 0.121 | 0.122 | 0.106 | 0.118 | 0.102 | 0.118 | 0.127 | 0.108 |
| \( \alpha = 0.05 \) | 0.058 | 0.070 | 0.061 | 0.054 | 0.067 | 0.051 | 0.065 | 0.073 | 0.056 |
| \( \alpha = 0.01 \) | 0.013 | 0.018 | 0.014 | 0.014 | 0.019 | 0.012 | 0.016 | 0.023 | 0.011 |

| **Swiss consumption survey – 1-factor model** | | | | | | | | |
| df | 74 | 128 | 194 | 74 | 128 | 194 | 74 | 128 | 194 |
| mean | 75.4 | 130.1 | 197.7 | 74.6 | 128.5 | 195.1 | 74.5 | 128.0 | 194.4 |
| \( \alpha = 0.20 \) | 0.229 | 0.239 | 0.254 | 0.214 | 0.209 | 0.221 | 0.210 | 0.202 | 0.207 |
| \( \alpha = 0.10 \) | 0.121 | 0.135 | 0.147 | 0.111 | 0.104 | 0.113 | 0.105 | 0.099 | 0.103 |
| \( \alpha = 0.05 \) | 0.067 | 0.076 | 0.086 | 0.056 | 0.055 | 0.060 | 0.051 | 0.053 | 0.053 |
| \( \alpha = 0.01 \) | 0.016 | 0.024 | 0.030 | 0.011 | 0.013 | 0.013 | 0.012 | 0.011 | 0.012 |

| **General social survey – 2-factor model** | | | | | | | | |
| df | 154 | 232 | 322 | 65 | 119 | 185 | 65 | 119 | 185 |
| mean | 154.8 | 234.0 | 323.3 | 65.6 | 119.7 | 185.5 | 0.217 | 0.215 | 0.217 |
| \( \alpha = 0.20 \) | 0.217 | 0.234 | 0.214 | 0.217 | 0.215 | 0.217 |
| \( \alpha = 0.10 \) | 0.113 | 0.131 | 0.116 | 0.114 | 0.111 | 0.113 |
| \( \alpha = 0.05 \) | 0.065 | 0.075 | 0.059 | 0.060 | 0.057 | 0.060 |
| \( \alpha = 0.01 \) | 0.018 | 0.022 | 0.018 | 0.013 | 0.013 | 0.017 |

Table 7 presents the number of times that the true bivariate parametric copulas are chosen over 100 simulation runs. If the true copula has distinct dependence properties with medium to strong dependence, then the algorithm performs extremely well as the sample size increases. Low selection rates occur if the true copulas have low dependence or similar tail dependence properties, since for that case it is difficult to distinguish
Table 7: Frequencies of the true bivariate copula identified using the model selection algorithm from 100 simulation runs.

| Political-economic dataset – 1-factor model | Continuous | Ordinal | 1-reflected Joe | Joe | reflected Joe | Joe | Gumbel |
|--------------------------------------------|------------|---------|----------------|-----|--------------|-----|--------|
| $n$                                        |            |         |                |     |               |     |        |
| 100                                        | 88         | 81      | 45             | 82  | 34           |
| 300                                        | 88         | 93      | 54             | 83  | 60           |
| 500                                        | 91         | 100     | 66             | 100 | 79           |

| General social survey – 1-factor model | Continuous | Ordinal | Count |
|---------------------------------------|------------|---------|-------|
| $n$                                   |            |         |       |
| 100                                   | 68         | 63      | 27    | 19 | 27  | 56  | 28    |
| 300                                   | 89         | 79      | 41    | 43 | 49  | 65  | 55    |
| 500                                   | 91         | 85      | 61    | 65 | 74  | 73  | 68    |

| Swiss consumption survey – 1-factor model | Continuous | Ordinal | Count |
|-------------------------------------------|------------|---------|-------|
| $n$                                       |            |         |       |
| 100                                       | 27         | 94      | 91    | 61 | 60  | 41  | 56    |
| 300                                       | 50         | 99      | 98    | 64 | 71  | 63  | 68    |
| 500                                       | 70         | 98      | 98    | 68 | 74  | 71  | 72    |

| General social survey – 2-factor model | Continuous | Ordinal | Count |
|---------------------------------------|------------|---------|-------|
| 1st Factor                             |            |         |       |
| $n$                                    |            |         |       |
| 100                                    | 22         | 40      | 10    | 19 | 19  | 50  | 6     |
| 300                                    | 26         | 52      | 11    | 42 | 36  | 79  | 16    |
| 500                                    | 19         | 67      | 13    | 52 | 53  | 83  | 39    |

| 2nd Factor                             | Continuous | Ordinal | Count |
|----------------------------------------|------------|---------|-------|
| $n$                                    |            |         |       |
| 100                                    | 13         | 28      | 28    | 7  | 14  | 21  | 17    |
| 300                                    | 26         | 39      | 56    | 30 | 45  | 28  | 47    |
| 500                                    | 32         | 67      | 65    | 53 | 59  | 33  | 70    |

| Swiss consumption survey – 2-factor model | Continuous | Ordinal | Count |
|------------------------------------------|------------|---------|-------|
| 1st Factor                               |            |         |       |
| $n$                                      |            |         |       |
| 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                               | Continuous | Ordinal | Count |
|------------------------------------------|------------|---------|-------|
| $n$                                      |            |         |       |
| 100                                      | 5          | 14      | 28    | 10 | 29  | 31  | 10    |
| 300                                      | 27         | 29      | 43    | 22 | 49  | 40  | 16    |
| 500                                      | 39         | 39      | 60    | 31 | 55  | 63  | 31    |

amongst parametric families of copulas (Nikolouloupolos and Karlis, 2008). For example,

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

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

8 Discussion

We have extended the factor copula models proposed in Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015) to the case of mixed continuous and discrete responses. They include the standard factor model as a special case and they can be seen to provide a substantial improvement over the latter on the basis of the log-likelihood principle, Vuong’s and $M_2$ statistics. Hence, superior statistical inference for the loading parameters of interest can be achieved.

This improvement relies on the fact that the latent variable distribution is expressed via factor copulas instead of the MVN distribution. The latter is restricted to linear and reflection symmetric dependence. Rizopoulos and Moustaki (2008) stressed that the inadequacy of normally distributed latent variables can be caused by the non-linear dependence on the latent variables. The factor copula can provide flexible reflection asymmetric tail and non-linear dependence as it is a truncated canonical vine copula (Brechmann et al., 2012) rooted at the latent variables. Joe et al. (2010) show that in order for a vine copula to have (tail) dependence for all bivariate margins, it is only necessary for the bivariate copulas in level 1 to have (tail) dependence and it is not necessary for the conditional bivariate copulas in levels $2, \ldots, d - 1$ to have tail dependence. The 1-factor copula has bivariate copulas with tail dependence in the 1st level and independence copulas in all the remaining levels of the vine (truncated after the 1st level). The 2-factor copula has bivariate copulas with tail dependence in the 1st and 2nd level and independence copulas in all the remaining levels (truncated after the 2nd level).

Even in the 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 $\tau$ (loading) parameters, the tail dependence varies, as explained in Section 2.1, and is a property to consider when choosing amongst different families of copulas and hence affects prediction. Rabe-Hesketh et al. (2003) highlighted the importance of the correct distributional assumptions for the prediction of latent scores. The estimated density of latent scores is simply the estimated density of latent variables. The latent scores will essentially show the effect of different model assumptions.
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).

because it is an inference that depends on the joint distribution. Figure 3 demonstrates these effects by revisiting the political-economic dataset 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). Note that even between the factor copula model with BVN copulas and the factor analysis model of Quinn (2004), there are small to moderate changes, because while these models share the same latent variables distribution, the former model does not assume the observed variables to be normally distributed, but rather uses the empirical distribution of the continuous observed variables, i.e. allows the margins to be quite free and not restricted by normal distribution. The differences at the lower panel graph are solely due the miss-specification the latent variable distribution.

As stated by many researchers (e.g., Rabe-Hesketh and Skrondal 2001, 2004), the major difficulty of all
the models with latent variables is identifiability. This issue as acknowledged by Hastie et al. (2001, page 494) has left many analysts skeptical of factor models, and may account for its lack of popularity in contemporary statistics. For example, for the standard factor model or the more flexible model in Irincheeva et al. (2012b) one of loadings in the second factor has to be set to zero, because the model with \(2d\) loadings is not identifiable. The standard factor model arises as special case of our model if we use as bivariate linking copulas the BVN copulas. Hence, for the 2-factor copula model with BVN copulas, one of the BVN copulas in the second factor has to be set as an independence copula. However, using other than BVN copulas, the 2-factor copula model is near-identifiable with \(2d\) bivariate linking copulas as it has been shown by Krupskii and Joe (2013) and Nikoloulopoulos and Joe (2015).

We tackled issues that particularly interest the social data analyst as model selection. Model selection in previous papers on factor copula models was mainly based on simple diagnostics that are informative if the variables are of the same type, e.g. item response (ordinal) data in Nikoloulopoulos and Joe (2015) or financial data in Krupskii and Joe (2013). This is not the case for mixed continuous and discrete data as the continuous data might have reflection asymmetric tail dependence, while the extreme-value behavior of discrete distributions is often degenerate (Feidt et al., 2010). As regard as to the issue of goodness-of-fit testing, we proposed a technique that is based on the \(M_2\) goodness-of-fit statistic (Maydeu-Olivares and Joe, 2006) in multidimensional contingency tables to overcome the shortage of goodness-of-fit statistics for mixed continuous and discrete response data (e.g., Moustaki and Knott 2000).

**Acknowledgements**

We would like to thank Professor Harry Joe (University of British Columbia) for comments leading to an improved presentation and Dr Irina Irincheeva (University of Bern) and Professor Marc Genton (King Abdullah University of Science and Technology) for providing the Swiss consumption survey dataset. 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.

**References**

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

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

Feidt, A., Genest, C., and Nešlehová, J. (2010). Asymptotics of joint maxima for discontinuous random variables. *Extremes*, 13:35–53.
Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer.

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

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

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

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

Irincheeva, I., Cantoni, E., and Genton, M. (2012a). A non-gaussian spatial generalized linear latent variable model. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(3):332–353.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

Nikoloulopoulos, A. K. and Karlis, D. (2008). Copula model evaluation based on parametric bootstrap. *Com-
Panagiotelis, A., Czado, C., Joe, H., and Stöber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis*, 106:138–152.

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

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

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

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

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

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

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

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

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

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

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

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

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

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