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

Aristidis K. Nikoloulopoulos

School of Computing Sciences
University of East Anglia

Email: `a.nikoloulopoulos@uea.ac.uk`

Joint work with Sayed H. Kadhem

September 30, 2021

Introduction

- It is very common in surveys, to deal with datasets with large number of items (ordinal variables) that are naturally divided into subgroups, in such, each group of items has homogeneous dependence.
- Factor models are a unified tool for the analysis of such datasets with dependence coming from a few latent variables/factors.
- In the literature, two factor models have been considered:
  - the bi-factor model (e.g., Gibbons and Hedeker 1992):
    - It consists of a common factor that is linked to all items, and non-overlapping group-specific factors.
    - The items are assumed to be (conditionally) independent given the group-specific and common factors.
  - the second-order model (e.g., de la Torre and Song 2009):
    - Items are indirectly mapped to an overall (second-order) factor via non-overlapping group-specific (first-order) factors.
    - The group specific (first-order) factors are linked to another (second-order) factor via an 1-factor model.

Motivation

- The existing models assume that the underlying continuous random variables follow a multivariate normal (MVN) distribution, thus, they can provide poor fit if
  - Items have more probability in joint upper or lower tail than would be expected with a (discretized) MVN;
  - items can be thought of as discretization of latent random variables that are maxima/minima or mixtures of means instead of means.

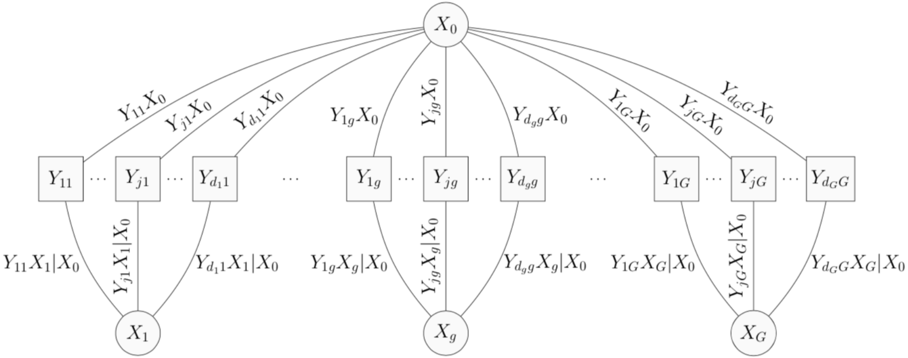
Solution: copula extensions for bi-factor and second-order

- The bi-factor copula model uses bivariate copulas to link the items to the common and group-specific factors.
- The second-order copula model uses bivariate copulas to link the first-order factors to the items and the second-order factor.
- They are truncated vine copula models (Brechmann et al., 2012) that involve both observed and latent variables.

Notation

- Let $\underbrace{Y_{11},\ldots,Y_{d_1 1}}_{1},\ldots,\underbrace{Y_{1g},\ldots,Y_{d_g g}}_{g},\ldots,\underbrace{Y_{1G},\ldots,Y_{d_G G}}_{G}$ denote the item response variables classified into the G non-overlapping groups.
- There are d_g items in group g; $g = 1,\ldots,G$, $j = 1,\ldots,d_g$.
- Collectively there are $d = \sum_{g=1}^{G} d_g$ items, which are all measured on an ordinal scale; $Y_{jg} \in \{0,\ldots,K_{jg}-1\}$.
- Let the cutpoints in the uniform $U(0,1)$ scale for the jg'th item be $a_{jg,k}$, $k = 1,\ldots,K_{jg}-1$, with $a_{jg,0} = 0$ and $a_{jg,K_{jg}} = 1$.
- These correspond to $a_{jg,k} = \Phi(\alpha_{jg,k})$, where $\alpha_{jg,k}$ are cutpoints in the normal $N(0,1)$ scale.

Bi-factor copula model



- Consider a common factor X_0 and G group-specific factors $X_1, \ldots, X_G$, where $X_0, X_1, \ldots, X_G$ are independent and standard uniformly distributed.
- $Y_{1g}, \ldots, Y_{d_g g}$ are conditionally independent given X_0 and X_g, and that Y_{jg} in group g does not depend on $X_{g'}$ for $g \neq g'$.

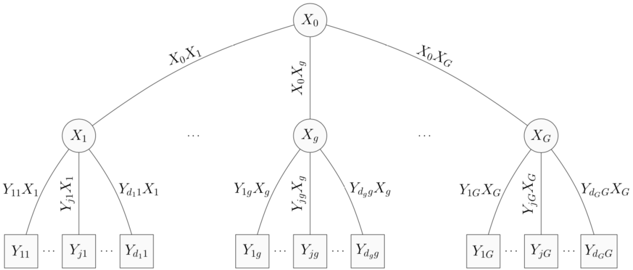
Joint probability mass function

- The joint probability mass function (pmf) is given by

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

- We specify $\Pr(Y_{jg}=y_{jg}|X_0=x_0,X_g=x_g)$ based on a set of bivariate copulas that link observed to latent variables:
  - According to Sklar's (1959) theorem there exists a bivariate copula C_{Y_{jg},X_0} such that $\Pr(Y_{jg}\le y_{jg},X_0\le x_0)=C_{Y_{jg},X_0}\big(F_{Y_{jg}}(y_{jg}),x_0\big)$, where C_{Y_{jg},X_0} is the copula that links observed variable with the common factor X_0.
  - Then it follows that $\Pr(Y_{jg}\le y_{jg}|X_0=x_0)=C_{Y_{jg}|X_0}\big(F_{Y_{jg}}(y_{jg})|x_0\big)$.
  - Then we let $C_{Y_{jg},X_g|X_0}$ be a bivariate copula that links the observed variable Y_{jg} with the group-specific factor X_g given X_0.
- Interestingly, the pmf reduces to an one-dimensional integral of a function which is in turn a product of G one-dimensional integrals.

Second-order copula model



- For a fixed $g = 1, \ldots, G$, the items $Y_{1g}, \ldots, Y_{d_g g}$ are conditionally independent given the first-order factors $X_g \sim U(0, 1)$, $g = 1, \ldots, G$.
- $X_1, \cdots, X_G$ are conditionally independent given the second-order factor $X_0 \sim U(0, 1)$.
- That is the joint distribution of $\mathbf{X} = (X_1, \cdots, X_G)$ has an one-factor structure.
- Y_{jg} in group g does not depend on $X_{g'}$ for $g \neq g'$.

Joint probability mass function

- The joint pmf takes the form

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

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

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

  where c_{X_g,X_0} is the bivariate copula density of the copula C_{X_g,X_0} linking X_g and X_0.
- Once again, specifying $\Pr(Y_{jg} = y_{jg} | X_g = x_g)$ based on a set of bivariate copulas, the pmf reduces to an one-dimensional integral of a function which is in turn a product of G one-dimensional integrals.

Bi-factor copula model

- If $C_{Y_{jg},X_0}(\cdot;\theta_{jg})$ and $C_{Y_{jg},X_g|X_0}(\cdot;\delta_{jg})$ are bivariate normal (BVN) copulas, then the Gaussian bi-factor model is a special case.
- The bi-factor copula model is the same as the Gaussian bi-factor model with stochastic representation

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

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

Second-order copula model

- For the Gaussian second-order model let $Z_0, Z'_1, \dots, Z'_G$ be the dependent latent $N(0,1)$ variables, where Z_0 is the second-order factor and $Z'_g = \beta_g Z_0 + (1-\beta_g^2) Z_g$ is the first-order factor for group g.

- For $g = 1, \dots, G$ and $j = 1, \cdots, d_g$, the stochastic representation is:

$$Z_{jg} = \beta_{jg} Z'_g + \sqrt{1-\beta_{jg}^2}\epsilon_{jg} \qquad Z'_g = \beta_g Z_0 + \sqrt{1-\beta_g^2} Z_g,$$

or

$$Z_{jg} = \beta_{jg}\beta_g Z_0 + \beta_{jg}\sqrt{1-\beta_g^2} Z_g + \sqrt{1-\beta_{jg}^2}\epsilon_{jg}.$$

- Hence, this is a special case of the Gaussian bi-factor model where $\theta_{jg} = \beta_{jg}\beta_g$ and $\gamma_{jg} = \beta_{jg}\sqrt{1-\beta_g^2}$.

IFM estimation

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

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

- We approach estimation using the two-step IFM method proposed by Joe (2005) that can efficiently, in the sense of computing time and asymptotic variance, estimate the model parameters.
  1. The univariate cutpoints for the jth item in group g are estimated as $\hat{a}_{jg,k} = \sum_{y=0}^{k} p_{jg,y}$, where $p_{jg,y}$, $y = 0, \ldots, K-1$ for $g = 1, \ldots, G$ and $j = 1, \ldots, d_g$ are the univariate sample proportions.
  2. The joint log-likelihood is maximized over the copula parameters with the cutpoints fixed as estimated at the first step.
- The estimated copula parameters can be obtained by using a quasi-Newton (Nash, 1990) method applied to the logarithm of the joint likelihood.

Bi-factor copula model

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

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

  2. Numerically evaluate the joint pmf

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

in a double sum

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

Second-order copula model

- For the second-order copula model numerical evaluation of the joint pmf can be achieved with the following steps:
  1. Calculate Gauss-Legendre quadrature points $\{x_q : q = 1, \ldots, n_q\}$ and weights $\{w_q : q = 1, \ldots, n_q\}$ in terms of standard uniform.
  2. Numerically evaluate the joint pmf

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

in a double sum

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

where $x_{q_2|q_1} = C^{-1}_{Y_{jg}|X_g;X_0}(x_{q_2}|x_{q_1};\delta_g)$.

- Note that the independent quadrature points have converted to dependent quadrature points that have an one-factor copula distribution $C_X(\cdot;\delta)$.

Bivariate copula selection

- In line with Nikoloulopoulos and Joe (2015), we use bivariate parametric copulas that can be used when considering latent maxima, minima or mixtures of means, namely the Gumbel, survival Gumbel (s.Gumbel) and Student t_ν copulas, respectively.
- We describe simple diagnostics based on semi-correlations and an heuristic method that automatically selects the bivariate parametric copula families that build either the bi-factor or the second-order copula model.
- In the context of items that can be split into G non-overlapping groups, such that there is homogeneous dependence within each group, it is sufficient to
  - summarize the average of the polychoric semi-correlations for all pairs within each of the G groups and for all pairs of items.
  - not mix bivariate copulas for a single factor; hence, for both the bi-factor and second-order copula models we allow $G+1$ different copula families, one for each group specific factor X_g and one for X_0.

Semi-correlations

- Choices of copulas with upper or lower tail dependence are better if the items have more probability in joint lower or upper tail than would be expected with the BVN copula.
- This can be shown with summaries of correlations in the upper joint tail and lower joint tail.
- Consider the underlying $N(0,1)$ latent variables Z_{jg}'s of the ordinal variables Y_{jg}'s.
- The correlations ρ_N^- and ρ_N^+ of Z_{jg}'s in the lower and upper tail, hereafter semi-correlations, depend only on the copula C of $\Big(\Phi(Z_{j_1g}),\Phi(Z_{j_2g})\Big)$; see (Joe, 2014, page 71).
- For the BVN and t_ν copulas $\rho_N^- = \rho_N^+$, while for the Gumbel and s.Gumbel copulas $\rho_N^- < \rho_N^+$ and $\rho_N^- > \rho_N^+$, respectively.
- The sample versions of ρ_N^+,ρ_N^- for item response data are the polychoric correlations in the joint lower and upper quadrants of Y_{j_1g} and Y_{j_2g} (Kadhem and Nikoloulopoulos, 2021).

Heuristic algorithm

1. Fit the bi-factor or second-order copula model with BVN copulas.
2. Fit all the possible bi-factor or second-order copula models, iterating over all the copula candidates that link all items Y_{jg}'s in group g or each group-specific factor X_g, respectively, to X_0.
3. Select the copula family that corresponds to the lowest Akaike information criterion (AIC), that is, $\text{AIC} = -2 \times \ell + 2 \times \#\text{copula parameters}$.
4. Fix the selected copula family that links the observed (bi-factor model) or latent (second-order model) variables to X_0.
5. For $g = 1, \ldots, G$:
    1. Fit all the possible models, iterating over all the copula candidates that link all the items in group g to the group-specific factor X_g.
    2. Select the copula family that corresponds to the lowest AIC.
    3. Fix the selected linking copula family for all the items in group g with X_g.

Goodness-of-fit

- We will use the limited information M_2 statistic proposed by Maydeu-Olivares and Joe (2006) to evaluate the overall fit of the proposed bi-factor and second-order copula models.
- For our parametric models with parameter vector θ of dimension q, let $\pi_2(\theta) = \big(\dot{\pi}_1(\theta)^\top, \dot{\pi}_2(\theta)^\top\big)^\top$ be the column vector of the univariate and bivariate model-based marginal probabilities that do not include category 0 with sample counterpart $\mathbf{p}_2 = (\dot{\mathbf{p}}_1^\top, \dot{\mathbf{p}}_2^\top)^\top$.
- The total number of the univariate and bivariate residuals $\big(\mathbf{p}_2 - \pi_2(\hat{\theta})\big)^\top$ is

$$s = d(K-1) + \binom{d}{2}(K-1)^2,$$

where $d(K-1)$ is the dimension of the univariate residuals and $\binom{d}{2}(K-1)^2$ is the dimension of the bivariate residuals excluding category 0.

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

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

  with

  $$\mathbf{C}_2(\boldsymbol{\theta}) = \Xi_2^{-1} - \Xi_2^{-1}\Delta_2(\Delta_2^\top \Xi_2^{-1}\Delta_2)^{-1}\Delta_2^\top \Xi_2^{-1} = \Delta_2^{(c)}\big([\Delta_2^{(c)}]^\top \Xi_2 \Delta_2^{(c)}\big)^{-1}$$

  where $\Delta_2 = \partial\boldsymbol{\pi}_2(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top$ is an $s \times q$ matrix with the first order derivatives of the univariate and bivariate marginal probabilities with respect to the estimated model parameters, $\Delta_2^{(c)}$ is an $s \times (s - q)$ orthogonal complement to Δ_2, such that $[\Delta_2^{(c)}]^\top \Delta_2 = \mathbf{0}$, and Ξ_2 is the asymptotic $s \times s$ covariance matrix of $\sqrt{n}\big(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})\big)^\top$.
- The limited information statistic M_2 under the null hypothesis has an asymptotic distribution that is χ^2 with $s - q$ degrees of freedom when the estimate $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent.

Simulations

- An extensive simulation study is conducted to
  1. gauge the small-sample efficiency of the IFM estimation method and investigate the misspecification of the bivariate pair-copulas;
  2. examine the reliability of using the heuristic algorithm to select the true (simulated) bivariate linking copulas;
  3. study the small-sample performance of the M_2 statistic.
- We randomly generate 1,000 datasets with samples of size $n = 500$ or 1000 and $d = 16$ items, with $K = 3$ or $K = 5$ equally weighted categories, that are equally separated into $G = 4$ non-overlapping groups from the bi-factor and second-order copula model.
- In each simulated model, we use different linking copulas to cover different types of dependence.
- For more details about the simulations see Kadhem and Nikoloulopoulos (2021).

Conclusions from the simulations

- IFM with the true bi-factor or second-order model is highly efficient according to the simulated biases, SDs and RMSEs.

- The IFM estimates of τ's are not robust under copula misspecification and their biases increase when the assumed bivariate copula has tail dependence of opposite direction from the true bivariate copula.

- The model selection algorithm performs extremely well for various bi-factor and second-order copulas models with different choices of linking copulas as the number of categories K increases.

- For a small K dependence in the tails cannot be easily quantified.

- The observed levels of M_2 are close to the nominal α levels and remain accurate even for extremely sparse tables ($d = 16$ and $K = 5$).

Application

- The Toronto Alexithymia Scale is composed of $d = 20$ items that can be subdivided into $G = 3$ non-overlapping groups:
  1. $d_1 = 7$ items to assess difficulty identifying feelings (DIF).
  2. $d_2 = 5$ items to assess difficulty describing feelings (DDF).
  3. $d_3 = 8$ items to assess externally oriented thinking (EOT).
- We use a dataset of 1925 university students from the French-speaking region of Belgium (Briganti and Linkowski, 2020).
- They were asked to respond to each item using one of $K = 5$ categories from “1 = completely disagree” to “5 = completely agree".
- For these items, a respondent might be thinking about the average “sensation" of many past relevant events, leading to latent means.
- Since the sample is a mixture (male and female students) we can expect a priori that a bi-factor or second-order copula model with t_ν copulas might be plausible, as in this case the items can be considered as mixtures of discretized means.

Semi-correlations

- We summarize the averages of polychoric semi-correlations for all pairs within each group and for all pairs of items along with the theoretical semi-correlations under different choices of copulas.

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

- For the first group of items there is more probability in the joint lower tail suggesting s.Gumbel linking copulas to join each item in this group with the DIF factor.
- For the other groups of items or for the items overall there is more probability in the joint lower and upper tail suggesting t_ν linking copulas.

Comparing factor structures

- We fit the bi-factor and second-order models with the copulas selected by the heuristic algorithm.
- For comparison, we also fit their special cases.
- The fitted models are compared via the AIC and the Vuong's test to show if (a) the best fitted model according to the AICs provides better fit than the other fitted models and (b) a model with the selected copulas provides better fit than the one with BVN.

	1-factor		2-factor		Bi-factor		Second-order	
	BVN	Selected	BVN	Selected	BVN	Selected	BVN	Selected
AIC	107135.8	105504.0	106189.5	103893.5	105507.7	103200.9	105878.6	104133.7
Vuong's 95% CI [a]	(0.35,0.50)		(0.53,0.69)		(0.51,0.69)		(0.38,0.52)	
Vuong's 95% CI [b]	(0.93,1.13)	(0.55,0.67)	(0.69,0.88)	(0.13,0.23)	(0.51,0.69)		(0.61,0.80)	(0.21,0.29)
M_2	14723.8	9865.0	9195.7	7383.7	11664.7	6381.5	13547.1	7341.2
df	3020	3020	3001	3000	3000	3000	3017	3017
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

- The best fitted bi-factor copula model results when we use s.Gumbel for the DIF factor, t_3 for both the DDF and EOT factors and t_2 for the common factor (alexithymia).

Goodness-of-fit

- It is not so clear from the goodness-of-fit *p*-values that the response patterns are satisfactorily explained by using the linking copulas selected by the heuristic algorithm.
- This is not surprising since one should expect discrepancies between the postulated parametric model and the population probabilities, when the sample size or dimension is sufficiently large (Maydeu-Olivares and Joe, 2014).
- To further show that the fit has been improved we have calculated the maximum deviations of observed and model-based counts for each bivariate margin.

	1-factor		2-factor		Bi-factor		Second-order	
	BVN	Selected	BVN	Selected	BVN	Selected	BVN	Selected
Items in Group 1	71	63	71	60	69	55	70	61
Items in Group 2	112	98	113	83	77	48	84	55
Items in Group 3	87	74	81	52	80	45	82	53
All items	112	98	113	83	80	55	84	61

- Overall, the maximum discrepancies have been sufficiently reduced in the selected bi-factor model.

Discussion

- For item response that can be split into non-overlapping groups, we have proposed bi-factor and second-order copula models.
- Our copula constructions include the Gaussian bi-factor and second-order models as special cases.
- They can provide a substantial improvement over the latter based on AIC, Vuong's and goodness-of-fit statistics.
- Hence, superior statistical inference for the loading parameters of interest can be achieved.
- The improvement relies on the fact that there can be an interpretation of latent variables that can be maxima/minima or mixture of means instead of means.
- Our models do not restrict to a latent structure that is additive.
- The copula parameters are interpretable as dependence of an observed variable with the common factor, or conditional dependence of an item with the group-specific factor given the common factor.

References

- Kadhem, S. H. and Nikoloulopoulos, A. K. (2021a). Factor copula models for mixed data. *British Journal of Mathematical and Statistical Psychology*.
- Kadhem, S. H. and Nikoloulopoulos, A. K. (2021b). Bi-factor and second-order copula models for item response data. *ArXiv e-prints*, arXiv:2102.10660.
- Krupskii, P. and Joe, H. (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis*, 120:85–101.
- Krupskii, P. and Joe, H. (2015). Structured factor copula models: Theory, inference and computation. *Journal of Multivariate Analysis*, 138:53–73.
- Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71:713–732.
- Nikoloulopoulos, A. K. and Joe, H. (2015). Factor copula models for item response data. *Psychometrika*, 80(1):126–150.