Factor 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 Harry Joe

Banff: May 23, 2013

Introduction

- Factor models are a unified tool for the analysis of high-dimensional responses with dependence coming from latent variables so that the number of dependence parameters is $O(d)$.
- We construct factor models based on copula functions for item response variables, $Y_1, \ldots, Y_d$ for d items, where the items (questions) are measured in an ordinal scale; $Y_j \in \{0, \ldots, K-1\}$ for $j = 1, \ldots, d$.
- The p-factor model assumes that $Y_1, \ldots, Y_d$ are conditionally independent given latent variables $X_1, \ldots, X_p$, and hence the joint probability mass function (pmf) is:

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

where $F_{X_1,\ldots,X_p}$ is the distribution of the latent variable.

Motivation

- We will use a general copula construction, based on a set of bivariate copulas that link observed to latent variables, to specify $\Pr(Y_j = y_j | X_1 = x_1, \ldots, X_p = x_p)$ and arrive at a very general conditional independence or factor model.
- Discretized multivariate normal (MVN) models with p-factor correlation matrices, are special cases of our general construction when all the above bivariate linking copulas are bivariate normal.
- Other choices of copulas are better if
  1. Y_j's have more probability in joint upper and/or lower tail than would be expected with discretized MVN;
  2. Y_j's can be considered as discretized maxima/minima or mixtures of discretized means rather than discretized means.
- For such items multivariate extreme value, elliptical distributions and copula theory can be used to select suitable copulas that link observed to latent variables.

Science item response data

This data set comes from the Consumer Protection and Perceptions of Science and Technology section of the 1992 Euro-Barometer Survey (Karlheinz and Melich, 1992). The questions (items) asked are given below:

1. Science and technology are making our lives healthier and easier.
2. Scientific and technological research cannot play an important role in protecting the environment and repairing it.
3. The application of science and new technology will make work more interesting.
4. Thanks to science and technology, there will be more opportunities for the future generations.
5. New technology does not depend on basic scientific research.
6. Scientific and technological research do not play an important role in industrial development.
7. The benefits of science are greater than any harmful effect it may have.

All of the items were measured on a four-group scale with response categories “0=strongly disagree", “1=disagree to some extent", “2=agree to some extent" and “3=strongly agree".

1-factor copula model

- Let the cutpoints in the uniform $U(0,1)$ scale for the jth item/variable be $a_{j,k}$, $k = 1, \ldots, K-1$, with $a_{j,0} = 0$ and $a_{j,K} = 1$.
- Let X_1 be a latent variable, which we assumed to be standard uniform. From Sklar (1959), there is a bivariate copula $C_{X_1 j}$ such that $\Pr(X_1 \leq x, Y_j \leq y) = C_{X_1 j}(x, F_j(y))$ for $0 \leq x \leq 1$ where F_j is the cdf of Y_j.
- Then it follows that $F_{j|X_1}(y|x) := \Pr(Y_j \leq y | X_1 = x) = \frac{\partial C_{X_1 j}(x, F_j(y))}{\partial x}$; let $C_{j|X_1}(a|x) = \partial C_{X_1 j}(x,a)/\partial x$ for shorthand notation.
- The pmf for the 1-factor model is

$$\begin{aligned} \pi_d(\mathbf{y}) &= \int_0^1 \prod_{j=1}^{d} \Pr(Y_j = y_j | X_1 = x)\, dx \\ &= \int_0^1 \prod_{j=1}^{d} \left[C_{j|X_1}(a_{j,y_j+1}|x) - C_{j|X_1}(a_{j,y_j}|x) \right] dx \end{aligned}$$

2-factor copula model

- Consider two latent independent uniform $U(0,1)$ variables X_1, X_2.
- 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,

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

- Here we are making the simplifying assumption that the conditional copula for the univariate distributions $F_{X_2|X_1} = F_{X_2}$ and $F_{j|X_1}$ does not depend on x_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 pmf for the 2-factor model is

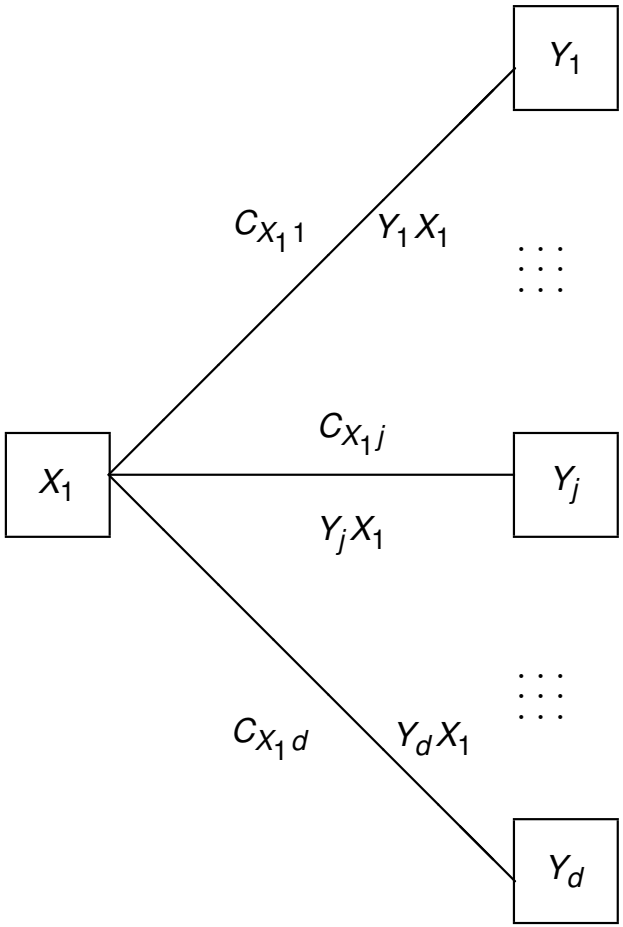
$$\pi_d(\mathbf{y}) = \int_0^1 \int_0^1 \prod_{j=1}^{d} \Pr(Y_j = y_j | X_1 = x_1, X_2 = x_2)\, dx_1\, dx_2$$

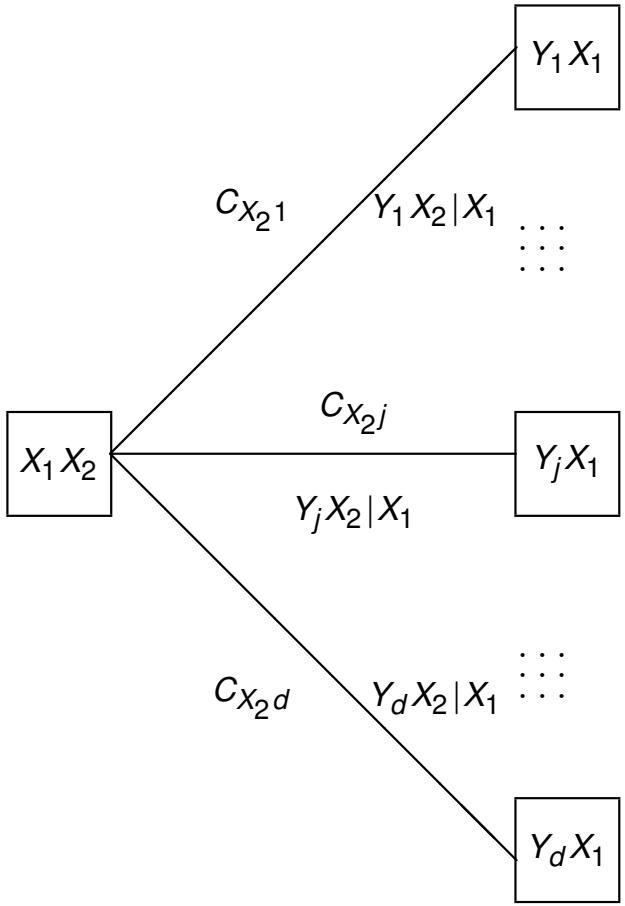
$$= \int_0^1 \int_0^1 \prod_{j=1}^{d} \left[C_{j|X_2}(F_{j|X_1}(y_j|x_1)|x_2) - C_{j|X_2}(F_{j|X_1}(y_j - 1|x_1)|x_2) \right] dx_1\, dx_2.$$

- For parametric 1-factor and 2-factor models, we let $C_{X_1 j}$ and $C_{X_2 j}$ be parametric bivariate copulas, say with parameters θ_j and δ_j respectively.
- Our general statistical model allows for selection of $C_{X_1 j}$ and $C_{X_2 j}$ independently among a variety of parametric copula families, i.e., there are no constraints in the choices of parametric copulas $\{C_{X_1 j}, C_{X_2 j} : j = 1, \ldots, d\}$.

Relationship with vines

These factor models can be explained as truncated canonical vines rooted at the latent variables.





The special case of the discretized MVN

- The pmf for the 1-factor model becomes,

$$\int_0^1 \prod_{j=1}^{d} \left\{ \Phi\left(\frac{\alpha_{j,y+1} - \theta_j \Phi^{-1}(x)}{\sqrt{1-\theta_j^2}} \right) - \Phi\left(\frac{\alpha_{j,y} - \theta_j \Phi^{-1}(x)}{\sqrt{1-\theta_j^2}} \right) \right\} dx$$

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{d} \left\{ \Phi\left(\frac{\alpha_{j,y+1} - \theta_j z}{\sqrt{1-\theta_j^2}} \right) - \Phi\left(\frac{\alpha_{j,y} - \theta_j z}{\sqrt{1-\theta_j^2}} \right) \right\} \phi(z)\, dz.$$

- Hence this model is the same as a discretrized MVN model with a 1-factor correlation matrix $R = (r_{jk})$ with $r_{jk} = \theta_j \theta_k$ for $j \neq k$.
- Similarly it can be shown that our model is a discretrized MVN model with a 2-factor correlation matrix $R = (r_{jk})$ with $r_{jk} = \theta_j \theta_k + \delta_j (1-\theta_j^2)^{1/2} \delta_k (1-\theta_k^2)^{1/2}$ for $j \neq k$.
- Note that the copula factor model formulation of the parameters is best seen through partial correlations for the second (and subsequent) factor(s).

Choices of bivariate copulas

1. Normal The resulting model in this case is the same as discretized MVN with factor correlation structure and has latent (ordinal) variables that can be considered as (discretized) means.

2. Gumbel The latent (ordinal) variables can be considered as (discretized) maxima, and there is more probability in the joint upper tail.

3. Survival Gumbel The latent (ordinal) variables can be considered as (discretized) minima, and there is more probability in the joint lower tail.

4. Student t_ν The latent (ordinal) variables can be considered as mixtures of (discretized) means. A small value of ν, such as $1 \leq \nu \leq 5$, leads to a model with more probabilities in the joint upper and joint lower tails compared with the normal copula.

Estimation

- For a sample of size N with data $\mathbf{y}_1, \ldots, \mathbf{y}_N$, the joint log-likelihood of the factor copula model is,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \pi_d(\mathbf{y}_i; \boldsymbol{\theta}). \tag{1}$$

- The Inference Function of Margins (IFM) method in Joe (2005, JMVA), can efficiently estimate the model parameters.
- In the first step, the univariate cutpoints are estimated as:

$$\hat{a}_{j0} = p_{j0},\ \hat{a}_{j1} = p_{j0} + p_{j1},\ \ldots,\ \hat{a}_{j,K-1} = p_{j0} + p_{j1} + \ldots + p_{j,K-1},$$

where $p_{jy},\ y = 0, \ldots, K-1$, for $j = 1, \ldots, d$ are the univariate sample proportions, and in the second step the joint log-likelihood (1) is maximized over the copula parameter vector with the cutpoints fixed at the estimated values from the first step.

Behavior of the log-likelihood for the 2-factor model

- For the special case of the 2-factor normal, one of $C_{X_{2j}}$ can be set as independence copula without loss of generality, because the model with $2d$ parameters is not identifiable.
- What happens if other copulas such as Gumbel and t_ν are used for bivariate linking copulas? Is the model with $2d$ bivariate linking copulas still not identifiable?

Conclusions from comparing the asymptotic covariance matrices

- A model with $2d$ Gumbel or t_ν copulas with $\nu \leq 3$ is clearly identifiable and the parameters can be interpreted.
- For t_ν with larger values of ν, we can set one of the $C_{X_{2j}}$ to be an independence copula, i.e., use $2d-1$ copulas, in order to make the parameters interpretable.

Table: Diagnostics based on the fit of the bivariate normal, Gumbel, s.Gumbel, and t_5 copulas, at each of the pair of items, comparing observed versus model-based bivariate counts with an emphasis on the tails.

Y_3	Y_7	observed	normal	Gumbel	s.Gumbel	t_5
0	0	4	3	3	5	5
0	1	8	11	11	11	11
0	2	12	15	15	13	12
0	3	9	3	4	4	5
1	0	6	7	7	6	6
1	1	34	29	30	29	31
1	2	47	47	48	47	47
1	3	11	15	13	16	14
2	0	8	9	10	8	7
2	1	52	50	50	49	49
2	2	111	104	106	106	111
2	3	35	43	40	44	40
3	0	3	1	1	2	3
3	1	6	10	9	11	10
3	2	23	28	24	28	24
3	3	23	17	21	14	19

Table: Estimated parameters, SEs in Kendall's τ scale and log-likelihoods ℓ, along with M_2 statistics (Maydeu-Olivares and Joe, 2006, Pka).

2-factor	normal	Gumbel/t_2		t_2/Gumbel		t_3/Gumbel	
	Est.	Est.	SE	Est.	SE	Est.	SE
θ_1	0.32	0.27	0.05	0.22	0.07	0.24	0.06
θ_2	-0.03	0.36	0.05	-0.18	0.07	-0.17	0.07
θ_3	0.38	0.15	0.05	0.36	0.06	0.37	0.05
θ_4	0.58	0.28	0.06	0.47	0.07	0.50	0.07
θ_5	-0.03	0.36	0.06	-0.24	0.08	-0.22	0.08
θ_6	0.09	0.44	0.05	-0.08	0.08	-0.06	0.07
θ_7	0.34	0.21	0.05	0.32	0.06	0.32	0.06
δ_1	0.13	0.20	0.07	0.24	0.06	0.24	0.06
δ_2	0.46	-0.31	0.07	0.42	0.05	0.43	0.05
δ_3	-0.09	0.36	0.06	0.13	0.07	0.11	0.07
δ_4	-0.01	0.49	0.07	0.30	0.09	0.30	0.09
δ_5	0.49	-0.37	0.07	0.47	0.06	0.48	0.06
δ_6	0.44	-0.21	0.09	0.50	0.05	0.50	0.05
δ_7	0.02	0.30	0.06	0.19	0.06	0.17	0.06
ℓ	-2921.9	-2864.7		-2866.3		-2866.7	
M_2	296.8	169.9		169.9		175.0	
df	176	175		175		175	
p-value	< 0.001	0.59		0.59		0.49	

Discussion

- We have proposed factor or conditional independence models where we replace bivariate normal distributions, between observed and latent variables, with bivariate copulas.
- It is the most general conditional independence model with univariate parameters separated from dependence parameters and latent variables that don't have necessarily an additive latent structure.
- Our factor copula construction includes the classic factor model as a special case and can provide a substantial improvement on the latter based on log-likelihood and goodness-of-fit.
- This improvement relies on the fact that when we use appropriate bivariate copulas other than normal copulas in the construction, there is an interpretation of latent variables that can be maxima/minima or high/low quantiles instead of means.

Extensions and future research

- The discrete factor model in Nikoloulopoulos and Joe (2013, Pka) can also easily be extended to other types of discrete data and to inclusion of covariates.
- Another direction of future research is to extend our factor model to capture the residual dependence:
  - Braeken et al. (2007, Pka) and Braeken (2011, Pka) explored the use of Archimedean copulas or a mixture of the independence and comonotonicity copulas to capture the residual dependence of the Rasch model.
  - A more general approach makes use of truncated vine copula models to model the residual dependence with $O(d)$ dependence parameters for d items.