Handling with categorical data in factor analysis

A copula-based approach

Antonio Calcagnì, Gianmarco Altoè, Massimiliano Pastore

Department of Developmental and Social Psychology
University of Padova

EMPG Meeting 2018
Genova, July 31 - August 2

Università degli Studi di Padova

Studies with **multivariate data** often involve different types of variables (e.g., continuous, ordinal, nominal).

Psychology, and more generally social sciences, often work with **categorical ordered** or **unordered** variables. Examples include rating scores, gender, counts.

Working with categorical variables usually requires appropriate statistical models, such as Generalized Linear Models (**GLMs**) in the case of linear conditional models.

In multivariate data analysis, models for categorical data include Structural Equation Models (**SEM**), Confirmatory Factor Analysis (**CFA**), and Correspondence Analysis (**CA**).

Some technical tricks are usually adopted to do estimations with categorical variables:

- Latent Variable Approach (Muthen, 1983)
- Multistage estimation (e.g., ULS, WLS, DWLS)
- Tetrachoric or polychoric approximations of the sample correlation matrix

In multivariate data analysis, models for categorical data include Structural Equation Models (**SEM**), Confirmatory Factor Analysis (**CFA**), and Correspondence Analysis (**CA**).

Moreover,

- computing standard errors and test statistics need some corrections (e.g., Satorra-Bentler, Satterthwaite)
- inference with small samples may be distorted
- due to numerical approximations, analysis of large datasets may be prohibitive

To retain as much as possible of the original variable metrics, **copulas** can be used to model the dependencies in the multivariate data.

Sklar's theorem (1959): every multivariate probability distribution $F(Y_1, \ldots, Y_k)$ can be represented by its univariate marginal distributions $F_1(Y_1), \ldots, F_k(Y_k)$ and a copula $\mathcal{C}$:

$$F(Y_1, \ldots, Y_k) = \mathcal{C}\big(F_1(Y_1), \ldots, F_k(Y_k) \mid \xi\big)$$

The dependence structure of the random vector $(Y_1, \ldots, Y_k)$ can be modeled considering marginals and copula separately.

Notably, copulas can be used to **generate random samples** from joint multivariate distributions of the involved variables.

Several copulas are available (e.g., **Gaussian**, Archimedean) for many applications.

Expectations for copulas are often known or approximated via Monte Carlo integration.

Categorical data in factor models

Using copulas

Recently, a novel **gaussian copula factor model** has been proposed for categorical data.

Bayesian Gaussian Copula Factor Models for Mixed Data

Jared S. MURRAY, David B. DUNSON, Lawrence CARIN, and Joseph E. LUCAS

Gaussian factor models have proven widely useful for parsimoniously characterizing dependence in multivariate data. There is rich literature on their extension to mixed categorical and continuous variables, using latent Gaussian variables or through generalized latent trait models accommodating measurements in the exponential family. However, when generalizing to non-Gaussian measured variables, the latent variables typically influence both the dependence structure and the form of the marginal distributions, complicating interpretation and introducing artifacts. To address this problem, we propose a novel class of Bayesian Gaussian copula factor models that decouple the latent factors from the marginal distributions. A semiparametric specification for the marginals based on the extended rank likelihood yields straightforward implementation and substantial computational gains. We provide new theoretical and empirical justifications for using this likelihood in Bayesian inference. We propose new default priors for the factor loadings and develop efficient parameter-expanded Gibbs sampling for posterior computation. The methods are evaluated through simulations and applied to a dataset in political science. The models in this article are implemented in the R package bfa (available from *http://stat.duke.edu/jsm38/software/bfa*). Supplementary materials for this article are available online.

KEY WORDS: Extended rank likelihood; Factor analysis; High dimensional; Latent variables; Parameter expansion; Semiparametric.

Recently, a novel **gaussian copula factor model** has been proposed for categorical data.

Interestingly, the model:

- adopts a **gaussian copula** to represent the dependence structure of the data
- works with both categorical and continuous variables in the same time (**mixed data**)
- is developed under the **Bayesian framework**
- can address many research questions via analysis of **posterior distributions**

Gaussian copula factor model

A brief review

In the ***standard* gaussian factor model** with J variables and K latent factors, we usually set:

$$
\begin{aligned}
\boldsymbol{\eta}^{(i)}_{K\times 1} &\sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_{K\times K}) \\
\boldsymbol{\epsilon}^{(i)}_{J\times 1} &\sim \mathcal{N}(\mathbf{0}_J, \boldsymbol{\Sigma}_{J\times J}) \\
\boldsymbol{y}^{(i)}_{J\times 1} &= \boldsymbol{\Lambda}_{J\times K} \cdot \boldsymbol{\eta}^{(i)}_{K\times 1} + \boldsymbol{\epsilon}^{(i)}_{J\times 1}
\end{aligned}
$$

where $\boldsymbol{\Sigma}$ is a (possibly) diagonal matrix of residuals.

By marginalizing out $\boldsymbol{\eta}$ from the joint distribution $(\boldsymbol{\eta}, \mathbf{y})$, we get marginal distribution for the observations only:

$$\mathbf{y}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma})$$

which indicates that, generally, $\text{cov}(\mathbf{y}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma}$ is a function of latent variables.

Gaussian copula factor model

A brief review

In the **gaussian *copula* factor model** (Murray et al., 2013) with J variables and K latent factors, we instead set:

$$
\begin{aligned}
\boldsymbol{\eta}^{(i)}_{K\times 1} &\sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_{K\times K}) \\
\boldsymbol{z}^{(i)}_{J\times 1} &\sim \mathcal{N}(\boldsymbol{\Lambda}_{J\times K}\cdot \boldsymbol{\eta}^{(i)}_{K\times 1}, \boldsymbol{I}_{J\times J}) \\
y_{ij} &= \mathcal{F}^{-1}\left(\Phi\left(z_{ij}/g(\boldsymbol{\lambda}_j)\right)\right)
\end{aligned}
$$

where:

- Φ is the univariate standard normal cdf
- $\mathcal{F}^{-1}$ are inverse of the margins of the copula
- $g(\boldsymbol{\lambda}_j) = \tilde{\boldsymbol{\lambda}}_j = \boldsymbol{\lambda}_j/\sqrt{1+\mathbf{1}_K\boldsymbol{\lambda}_j^2}$ are **scaled loadings**

In the **gaussian *copula* factor model** (Murray et al., 2013) with J variables and K latent factors, we instead set:

$$
\begin{aligned}
\boldsymbol{\eta}^{(i)}_{K\times 1} &\sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_{K\times K}) \\
\boldsymbol{z}^{(i)}_{J\times 1} &\sim \mathcal{N}(\boldsymbol{\Lambda}_{J\times K}\cdot \boldsymbol{\eta}^{(i)}_{K\times 1}, \boldsymbol{I}_{J\times J}) \\
y_{ij} &= \mathcal{F}^{-1}\left(\Phi\left(z_{ij}/g(\boldsymbol{\lambda}_j)\right)\right)
\end{aligned}
$$

Note that:

- the correlation c between two variables j' and j'' is: $c_{j'j''} = \tilde{\boldsymbol{\lambda}}_{j'}^T \tilde{\boldsymbol{\lambda}}_{j''}$
- $\tilde{\boldsymbol{\Lambda}}$ governs the dependence structure separately from the marginal distributions, i.e. we are not decomposing $\text{cov}(\mathbf{y})$

The **gaussian *copula* factor model** (Murray et al., 2013) is identified by minimal conditions (sign constraints and fixed zeros in $\mathbf{\Lambda}$, fixed K).

Model parameters are represented in terms of (posterior) probability distributions via Paramater-Expanded (**PX**) **Gibbs Sampler** targeting on the joint posterior density $f(\tilde{\mathbf{\Lambda}}, \mathbf{N}|\mathbf{Y})$, with **N** being the matrix of entries $\boldsymbol{\eta}_i$.

Prior distribution over $\tilde{\mathbf{\Lambda}}$ is coniugate (Murray et al., 2013): **Generalized (double) Pareto**.

An illustrative example

Assess attachment in children with ECR-RC

Measures: 12 (five-point Likert scale) items of ECR-RC, a short questionnaire to assess anxious and avoidant attachments in children and adolescents (Brenning, 2015).

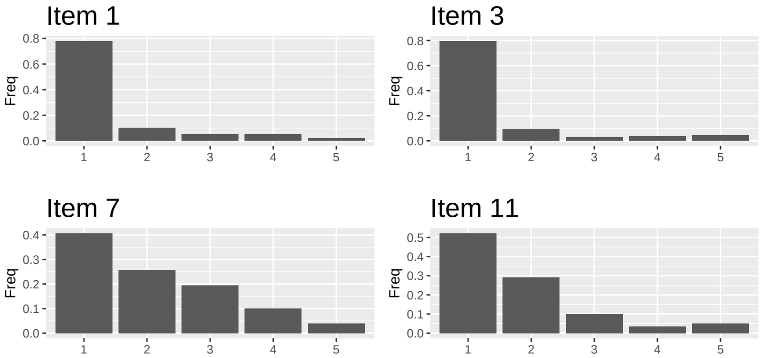
Sample: 259 Italian children (51% girls), mean age = 4 years and 2 months, SD = 7 months, range = 8.2 - 10.3

Factor structure: two latent factors, *anxiety* and *avoidance*.

An illustrative example

Assess attachment in children with ECR-RC

Variables are represented as ordered categories. Here, some item response distributions:



An illustrative example

Assess attachment in children with ECR-RC

We followed Marci et al. (2018) and defined a factorial model with 12 items and 2 latent factors.

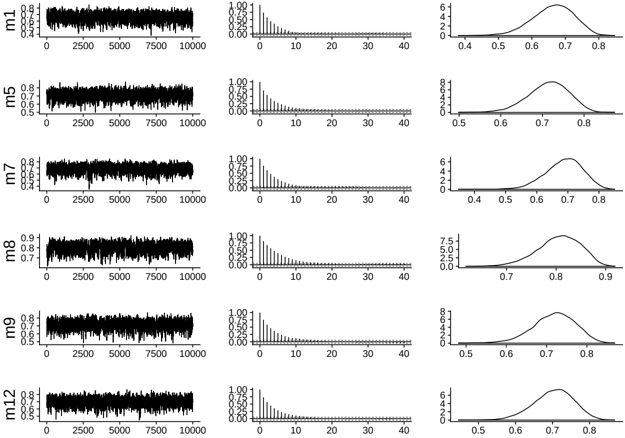
The model was fit using the R package `bfa` (Murray, 2016).

Variables in the data frame were re-coded as ordered factors, priors on loadings were modeled as GDP (default choice), MCMC-samples = 10000, initial burnin = 2500.

An illustrative example

MCMC Results

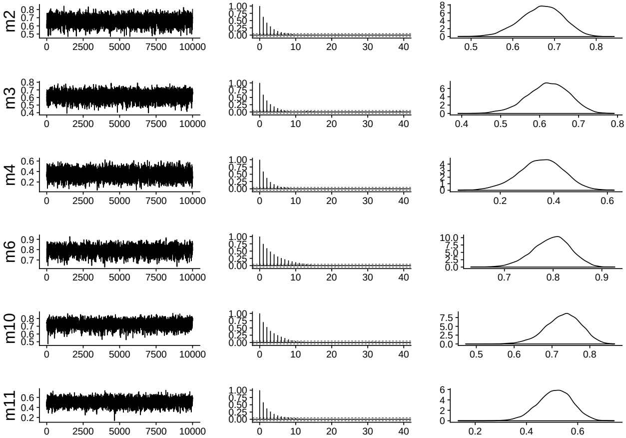
Factor: *Anxiety*



An illustrative example

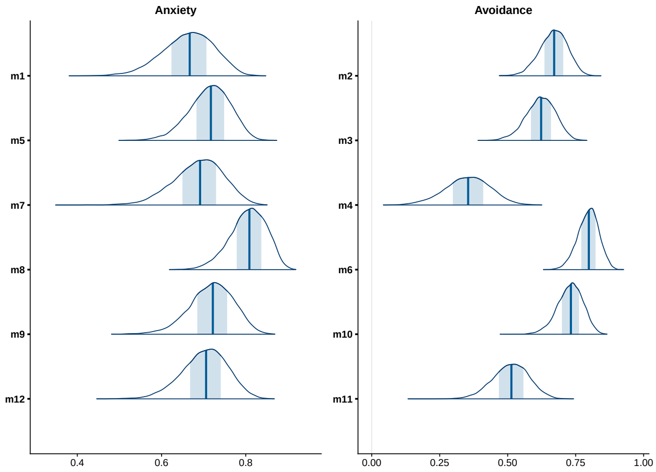
MCMC Results

Factor: *Avoidance*



An illustrative example

Posterior Results



Conclusions

Potentials

- The model **adequately works** with ordered and unordered categorical data
- The Bayesian framework:
  - **overcomes many limits** of the standard LS or ML estimation approaches
  - offers a way to do (posterior) **data analysis** in this type of models
- This approach allows a great deal of **flexibility** in analysing skewed and non-gaussian variables while modeling the multivariate dependencies

Conclusions

Limits and developments

- The **model lacks** a way to model:
  - the covariances among latent variables cov(η)
  - the errors of the measurements model

- This approach works like a “smart” principal component analysis where constraints can be set in the latent structure

Further developments will consider:

- testing the model over a detailed simulation scenario
- extending the model to modeling covariances among latent factors and measurement errors

antonio.calcagni@unipd.it
gianmarco.altoe@unipd.it
massimiliano.pastore@unipd.it

LaTeX

useR!

knitr

Elegant, flexible and fast dynamic report generation with R

Stan