T E C H N I C A L
R E P O R T

0327

A NOTE ON A COPULA APPROACH TO POLYCHORIC CORRELATIONS

C. ALMEIDA and M.MOUCHART

I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Note on a copula approach to polychoric correlations*

Carlos Almeida AND Michel Mouchart†

Institut de Statistique
Université Catholique de Louvain
20, Voie du Roman Pays
1348 Louvain-la-Neuve
Belgium

Abstract

Polychoric correlations among ordinal variables rest on the interpretation of the ordinal variables as discretization of latent continuous "ideally measured" variables and the assumption that these corresponding latent variables are jointly normally distributed. As the marginal distribution of these latent variables are arbitrary, for not being identified, the copula concept is a natural tool for the specification of the dependence structure. The role of the multivariate normality hypothesis is reexamined from a copulistic specification.

Keywords: Ordered Variables, Polychoric Correlations, Identification, Gaussian multidimensional copula.

*Financial support from the contract 'Projet d'Actions de Recherche Concertées' nr 98/03–217 from the Belgian government, and from the IAP research network nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged. Without implications, the authors gratefully acknowledge useful comments on earlier versions by Ph. Lambert.

†Corresponding author: mouchart@stat.ucl.ac.be

When dealing with polychoric correlations, see *e.g.* Olsson (1979), Jöreskog (1994), the main hypothesis is that for each manifest $X_k \in \{1, \dots r_k\}$, $k = 1, \dots, K$, there is a continuous latent random variable (*ideally measured variable*) $X_k^* \in \mathbb{R}$, and an array

$$\mathcal{A} = \{a_j^{(k)} : k \in \{1, \dots, K\}, \quad j \in \{1, \dots, r_k - 1\}\}, \tag{1}$$

with $-\infty = a_0^{(k)} < a_1^{(k)} < \dots < a_{r_k}^{(k)} = \infty$, such that:

$$\forall\ j = \{1, \dots, r_k\} \qquad X_k \leq j \Leftrightarrow X_k^* \leq a_j^{(k)}. \tag{2}$$

Denoting $X = (X_1, \dots, X_K)$, $X^* = (X_1^*, \dots, X_K^*)$ and $\underline{j} \in R_X = \displaystyle\prod_{k=1}^{K} \{1, \dots, r_k\}$, we have:

$$\forall\ \underline{j} \in R_X : \mathbb{P}_{\theta_{DL}}(\ X \leq \underline{j}\ ) = \mathbb{P}(\ \forall\, k \in \{1, \dots, K\},\ X_k^* \leq a_{j_k}^{(k)}\ ), \tag{3}$$

where DL stands for discrete latent and " $\leq$ " in the left side of (3) is the coordinate-wise order. The model is eventually parametrized by:

$$\theta_{DL} = (F, \mathcal{A}), \tag{4}$$

where F is the multivariate probability distribution of the ideally measured variables X^* and $\mathcal{A}$ gathers the thresholds defining the discretization. Thus the array $\mathcal{A}$ operates a decomposition of $\mathbb{R}^K$ into $\prod_{k=1}^{K} r_k$ cubes, namely: $c_{\underline{j}} = c_{j_1, \dots, j_K} = \displaystyle\prod_{k=1}^{K} (a_{j_k - 1}^{(k)}, a_{j_k}^{(k)}]$, and:

$$\mathbb{P}_{\theta_{DL}}(\ X = \underline{j}\ ) = F(c_{\underline{j}}). \tag{5}$$

As this approach is invariant under strictly increasing coordinate-wise transformations of X_i^*, this property leaves the marginal univariate distributions of these ideally measured variables unidentified (for detail, see *e.g.* Almeida and Mouchart (2003)).

Now, the *matrix of polychoric correlations* is defined as the $k \times k$ correlation matrix of the continuous real valued random variables $\{X_k^* : k = 1, \dots, K\}$, *viz.* $\Gamma = (\gamma_{ij})$, where $\gamma_{ij} = \text{Corr}\ (X_i^*, X_j^*)$. A serious drawback of this concept is not to be invariant under strictly increasing (non linear) transformations of X_i^*. The concept of polychoric correlations is most natural under the normality hypothesis in which case (5) becomes:

$$\mathbb{P}_{\theta_{NPO}}(\ X = \underline{j}\ ) = \Phi_\Gamma(c_{\underline{j}}) \text{ with } \theta_{NPO} = (\Gamma, \mathcal{A}), \tag{6}$$

where NPO stands for normal polychoric and where Φ_Γ is the multivariate normal distribution with zero mean, unit variance and Γ correlation matrix.

The *identifiability of this model* is provided by the following result, the proof of which may be found in Almeida and Mouchart (2003)

Theorem 1. **(i)** *Under the normality hypothesis, $\theta_{NPO} = (\Gamma, \mathcal{A})$ is identified, or equivalently, the mapping $\theta_{NPO} \mapsto \mathbb{P}_{\theta_{NPO}}$, defined in (6), is one-to-one, if the polychoric correlations matrix Γ is not singular.*

(ii) *Under the hypothesis of theorem 1, θ_{NPO} is just identified if $K = r_1 = r_2 = 2$; otherwise, the model is overidentified.*

■

As the marginal distributions of the ideally measured variables are not identified, it is natural to decompose the characterization of the multivariate distribution of these variables into, on the one hand, the set of its marginal distributions and, on the other hand, a residual aspect that would be variation-free with respect to the first one and would capture the association between the latent variables. This is exactly the idea of a *copula*, which is a multivariate distribution function with uniform margins on the unit interval. A detailed study of copulas may be found in Nelsen (1999), in particular the following basic result.

Theorem 2. **(i)** *Let $F_1, \ldots, F_K$ be K univariate distributions and C a K-dimensional copula. Then $F(x_1, \ldots, x_K) = C(F_1(x_1), \ldots, F_K(x_K))$ defines a multivariate distribution F with margins $F_1, \ldots, F_K$.*

(ii) *(Sklar's theorem) Let F be a K-dimensional distribution with continuous margins $F_1, \ldots, F_K$. Then F has a unique copula representation*

$$F(x_1, \ldots, x_K) = C(F_1(x_1), \ldots, F_K(x_K)). \tag{7}$$

■

Denoting by Φ the standard normal univariate distribution function, the *gaussian copula* corresponding to any correlation matrix Γ is defined by:

$$C_\Gamma^G(u_1, \ldots, u_K) = \Phi_\Gamma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_K)) \quad (u_1, \ldots, u_K) \in [0,1]^K$$

From Sklar's theorem, there is a unique copula C satisfying (7). The discretization model (2)-(3) can therefore be parametrized, instead of (4), as follows:

$$\theta_{DLC} = (\{F_k : k = 1, \ldots, K\}, C, \mathcal{A}), \tag{8}$$

The non-identified margins F_k can be fixed arbitrarily to be $\mathcal{U}[0,1]$, in which case the corresponding threshold values are included in $[0,1]$; we stress this fact denoting by $p_j^{(k)}$ the thresholds corresponding to a copula. The model (2)-(3) may now be rewritten as:

$$\forall\ \underline{j} \in R_X \quad X \leq \underline{j} \Leftrightarrow \{X_k^* \leq p_{j_k}^{(k)}, \quad k = 1, \ldots, K\}, \tag{9}$$

$$\mathbb{P}(\ X \leq \underline{j}\ ) = C(p_{j_k}^{(k)} : 1 \leq k \leq K). \tag{10}$$

Thus $\theta_{CO} = (C, \mathcal{P})$ is a sufficient parametrization, instead of (8), once uniform margins are specified for identification proposes. Therefore (1) is reparametrized into:

$$\mathcal{P} = \{p_j^{(k)} : j = 1, \ldots, r_k - 1,\ k = 1, \ldots, K\} \qquad p_j^{(k)} = F_k(a_j^{(k)}) \in [0, 1]$$

The thresholds $p_j^{(k)}$ admit two simple and natural interpretations: $p_j^{(k)}$ is the probability that the manifest (ordinal) variable takes a value equal than or inferior to j and is such that $a_j^{(k)}$ corresponds to the $p_j^{(k)}$-quantile of the unidentified marginal distribution F_k. Moreover, the estimation of the thresholds $p_j^{(k)}$ is direct using the empiric proportions and this estimation is unbiased and consistent. Note that the thresholds $a_j^{(k)}$ may be estimated consistently by $\hat{a}_j^{(k)} = F_k^{-1}(\hat{p}_j^{(k)})$ for any arbitrary F_k.

The gaussian copula version of model (2)-(3) becomes:

$$\mathbb{P}_{\Gamma,\mathcal{P}}(\ X \leq \underline{j}\ ) = C_\Gamma^G(p_{j_k}^{(k)} : 1 \leq k \leq K), \tag{11}$$

the parametrization of which now becomes $\theta_{NCO} = (\Gamma, \mathcal{P})$, that is clearly in bijection with θ_{NPO}. Therefore, conditions of application of theorem 1 remain the same.

As a conclusion, this note has stressed that the normality hypothesis underlying the concept of polychoric correlations actually bears on the gaussianity of the underlying copula: the normality (and standardization) of the margins is not identified and irrelevant. This fact leaves open the testing of the gaussianity assumption. Some packages propose to test bivariate normality pairwise because complete joint normality raises substantial operational difficulties, but there is some hope that a copula approach along with resampling methods could provide interesting prospects.

References

Almeida, C. and Mouchart, M. (2003). *Identification of polychoric correlations: A copula approach*, Discussion Paper, Institut de statistique, Université catholique de Louvain.

Jöreskog, K. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix, *Psychometrika* **59**, 381–389.

Nelsen, Roger B. (1999). *An introduction to copulas. Properties and applications.* Lecture Notes in Statistics (Springer) **139**.

Olsson, Ulf (1979). Maximum likelihood of the polychoric correlation coefficient, *Psychometrika* **44**, 443–460.