Factor Copula Analysis for Multivariate Ordinal Data

by

Agnes Nessie Amu

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics

University of Manitoba

Winnipeg

Abstract

Factor analysis is commonly used in many fields to establish relationships among manifest variables in terms of few underlying latent factors. Besides the elucidation of these factors, of importance in many applications is their subsequent analysis, for instance, in regression settings given some covariates. Motivated by genetic association studies of autism spectrum disorders, this research revisits common dependence measures for multivariate ordinal data, and investigates robustness of factor analysis and factor scores regression under various distributional settings.

We first demonstrate numerically, a comparison of dependence measures for multivariate ordinal data and investigate the robustness of polychoric correlation estimation under settings with asymmetric dependence patterns and varying degrees of skewness in marginal distributions. To accommodate such general joint distributions, we make use of factor copula models in data generation. These models offer a very general and flexible framework to study relationships among manifest and latent variables. Hence, we propose an alternative strategy to quantify the scores on the latent variables using factor copula scores and investigate the performance of the proposed approach in comparison to the traditional factor model both in factor scores estimation and their association testing for a given covariate. Our findings suggest that

the traditional factor analysis is considerably robust to violations of distributional assumptions. Factor copula analysis yields very similar results to those based on polychoric correlation, with a slight power gain in association testing.

Acknowledgment Page

Foremost, I want to thank the Almighty God for His grace and mercy over my life and for giving me strength to finish this program. My sincere gratitude goes to my advisor Dr. Elif Acar. Her motivation and encouragement boosted my confidence whenever I was discouraged. Her contributions to the direction of this thesis are innumerable. Thank you very much. Special thanks to my committee members; Dr. Lisa Lix and Dr. Saman Muthukumarana for their guidance and insightful comments. To my colleagues and friends, I say thanks for being there. Also, to my mother Elizabeth Mensah, my sister and brother, thanks for all the support and prayers. To my God father, Mr. Joseph Osei, God richly bless you for everything. To my husband, Emmanuel Oduro Asante thanks for the love and the emotional support.

Dedication Page

This thesis is dedicated to my mother. She sacrificed her dreams so that I could have my own.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many applications, variables of major importance are not directly observable and can only be assessed through some manifest variables. For instance, in psychometric research variables such as level of satisfaction or intelligence cannot be directly measured but can be quantified using opinion surveys or carefully designed questionnaires. Psychologists measure the concept of intelligence by developing a hypothesis of intelligence and write measurement instruments with items (questions) designed to assess various cognitive abilities (Salkind, 2007). Similarly, the assessment of many neurological and psychological disorders relies heavily on diagnostic questionnaires, which consist of a moderate to large number of likert scale questions. A suitable statistical analysis for such data must take into account both the scale and dimensionality of measurements.

Factor analysis is a multivariate technique to analyze and explain covariance relationships among random variables in terms of few underlying, but unobservable (latent) random quantities called *factors*. It is commonly used in areas such as psychometrics (Fabrigar et al., 1999), economics (Bai et al., 2015), physical and biological sciences (Kaplunovsky, 2005) among others. There are two types of factor analysis, classified as exploratory and confirmatory depending on the research interest. Exploratory factor analysis concerns identification and interpretation of latent factors, while confirmatory factor analysis is used to test relationships among latent and manifest variables. In this thesis we use the term "factor analysis" in broad terms, but mostly to refer to exploratory factor analysis.

While factor analysis is primarily used to elucidate latent factors from the dependence mechanisms in observed data, it also serves as a dimension reduction technique. The latter aspect is particularly appealing when interest lies in subsequent analysis of identified factors rather than a direct analysis of manifest variables. For instance, in genetic studies of complex psychological disorders, factor analysis is employed to identify latent quantitative traits describing unobserved disease features. These identified traits are then used in genetic testing of disease associated markers, usually performed using a regression-type model. See, for instance, Liu et al. (2011) and Bralten et al. (2017) which employ factor analysis to define quantitative traits of autism spectrum disorders (ASD) and Cavallini et al. (2002) of obsessive compulsive

disorder (OCD). We also refer the reader to Preacher and MacCallum (2002) and Brown et al. (2015) for other genetic applications of factor analysis.

Factor analysis methods typically require manifest variables to be quantitative and jointly normally distributed. When these assumptions are satisfied, factor analysis is performed using the Pearson correlation matrix of the observed data, and fitting a factor model under the likelihood framework for the multivariate normal distribution. An overview of factor analysis and commonly used estimation procedures can be found in standard multivariate analysis textbooks (e.g., Johnson et al., 2014).

However, in most psychometric applications, manifest variables are ordinal in nature, hence violate the distributional assumptions of factor analysis. In such cases, a common practice is to replace the Pearson correlation matrix with the polychoric correlation matrix of the multivariate ordinal data, and keep the multivariate normal distribution framework for fitting a factor model. This approach relies on an implicit assumption that the ordinal variables arise from a discretization of multivariate normal latent variables. Such distributional assumptions are often adopted for convenience in model fitting, but are questionable in practice due to latency of underlying variables.

A number of studies have investigated robustness of factor analysis under violations of the normality assumption. Jin and Yang-Wallentin (2017) found that the polychoric correlation can be underestimated when the underlying distribution is misspecified as normal. They concluded that, the bias in the

estimates can be higher when the true underlying distribution is skewed. The results from Olsson (1979) demonstrate that for highly skewed variables, the factor analysis produces biased loadings and a lack of fit of the model for the observed variables. These investigations mostly focus on evaluation of polychoric correlation estimates and consider non-normality in the marginal distributions, i.e. through different degrees of skewness and kurtosis. While polychoric correlation estimates are concluded to be fairly robust against violations of the normality assumption, similar assessments for factor scores estimates and their subsequent analysis have not been studied in detail. Furthermore, violations of normality in joint distributions (e.g., non-elliptical dependence structures) have not been considered. This was partially due to the lack of a rich class of latent variable models for multivariate ordinal data generation.

There have been some efforts to extend factor models to allow non-normal joint distributions. Klüppelberg and Kuhn (2009) introduced a copula factor model for (meta-)elliptical distributions (e.g., multivariate t), which admits a correlation-based dependence structure. Krupskii and Joe (2013) and Krupskii (2014) proposed a more general framework which allows arbitrary bivariate linking copulas and non-elliptical dependence structures, and referred the resulting class of model as factor copula models. The focus in these work has been on quantitative manifest variables. More recently, Nikoloulopoulos and Joe (2015) extended factor copula models to multivariate ordinal data and discussed parameter estimation and model fitting. However, estimation of factor

scores from these models and their subsequent use have not been addressed.

This research revisits the dependence measures for multivariate ordinal data, and investigates robustness of factor analysis and factor scores regression under various distributional settings. Our motivation lies in genetic association studies of autism spectrum disorders (ASD), which typically consider the presence or absence of the disorder and do not utilize the multivariate ordinal data from diagnostic questionnaires. Since disease diagnosis is difficult and its severity can greatly vary, identification of latent quantitative traits through factor analysis offers better strategies to quantify the severity of the disorder and to assess its genetic architecture.

Here, our first objective is to provide a detailed comparison of dependence measures for multivariate ordinal data. We consider the Pearson- and Spearman-type correlation measures and investigate their robustness under settings with asymmetric dependence patterns and varying degrees of skewness in marginal distributions. To accommodate such general joint distributions, we utilize factor copula models in data generation. Secondly, we examine the robustness of polychoric correlation-based factor analysis against violations of the normality assumption. In our assessments we consider the estimates of both the model parameters and factor scores. Thirdly, we propose factor copula scores as an alternative strategy to quantify the latent traits. For this we use factor copula models, which provide a very general and flexible framework not only to generate data but also to estimate latent factors. Lastly, we

compare the performances of these approaches in factor scores regression, for testing genetic associations.

The remaining of this chapter provides an overview of ordinal data methods and copula models.

1.1 Ordinal Variables

Ordinal variables are categorical variables where the defined categories obey a natural order. These variables cannot be treated the same as nominal variables due to their ordered relation in the categories. They also differ from continuous variables in that the differences between the ordered categories may not be equal. Some examples of ordinal variables are student letter grades (e.g., A, B, C, etc.), likert scale responses in opinion polls (e.g., strongly disagree, disagree, agree, etc.), and income levels (e.g., low, medium and high). While the categorical attributes in ordinal data can be coded numerically, statistical methods should be tailored so as to take into account any trend or order in the aspects of interest.

The statistical analysis of ordinal variables include contingency table analysis, cumulative link models and latent variable models among others (Alan, 2010). Contingency tables provide a simple way to summarize the frequency distribution of categorical variables, for which the associations are tested using the Pearson chi-squared test or the Fisher exact test. While these

approaches can also be used for ordinal variables, statistical tests such as Cochran-Armitage trend test and Cochran-Mantel-Haenszel (CMH) test, that account for trends in relationships are often more suitable.

Another method of analyzing ordinal variables is the cumulative link models which uses log-odds ratio to test variable associations. While these models provide a general approach for regression-type settings, the interpretations of variable relationships are often not straightforward and can be quite complex. The latent variable models, on the other hand, define a class of models that treat ordinal variables as discretization of inherently continuous unobserved (latent) variables. In many applications, they are preferred over other approaches for their ease of interpretations. This is also the case in the current study.

Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ denote a d-dimensional vector of ordinal random variables and $\mathbf{Y}^* = (Y_1^*, \ldots, Y_d^*)$ be the corresponding latent variable vector, which are related through the relations,

$$\begin{aligned}
Y_i = 0 \quad &\Longleftrightarrow \quad Y_i^* < a_{1i} \\
Y_i = 1 \quad &\Longleftrightarrow \quad a_{1i} \leq Y_i^* < a_{2i} \\
&\vdots \qquad\qquad \vdots \\
Y_i = K_i \quad &\Longleftrightarrow \quad a_{Ki} \leq Y_i^*,
\end{aligned}$$

where $\{a_{1i}, \ldots, a_{Ki}\}$ is the set of unknown thresholds for each $i = 1, \ldots, d$. Figure 1.1 displays the relationship between Y_1 and Y_1^* in the case where $K = 3$ and Y_1^* has a standard normal distribution.

Figure 1.1: The relationship between Y_1 and Y_1^*.

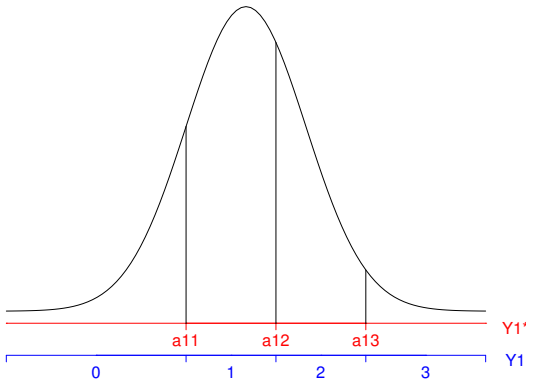
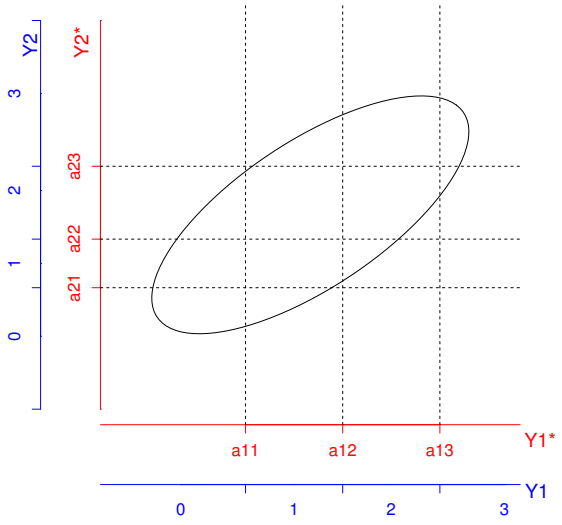


Figure 1.2: The pairwise relationship between Y_1 and Y_2, where Y_1^* and Y_2^* have the standard bivariate normal distribution with $\rho = 0.7$.



In latent variable models, the dependence among the observed (manifest) variables are dictated by the dependence among the latent variables. Hence, when studying dependence in $\mathbf{Y}$, the underlying distribution of $\mathbf{Y}$ plays a crucial role. Figure 1.2 illustrates the pairwise dependence between Y_1 and Y_2, each with $K = 3$ categories, through that of Y_1^* and Y_2^*, where the latter have the standard bivariate normal distribution with correlation $\rho = 0.7$

The relationship between $\mathbf{Y}$ and $\mathbf{Y}^*$ further facilitates the factor analysis of multivariate ordinal data through the polychoric correlation, which will be detailed in Chapter 3.

1.2 Copulas

Copula models are flexible multivariate models, which can accommodate continuous, discrete and mixed type data. In recent years, these models have gained increasing popularity in finance, insurance and hydrology among other fields.

Copulas are multivariate distribution functions with standard uniform margins. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be a random vector with the joint distribution F and marginal distribution functions F_i for $i = 1, \ldots, d$. Then, by Sklar's Theorem (Sklar, 1959),

$$F(y_1, \ldots, y_d) = C\left\{F_1(y_1), \ldots, F_d(y_d)\right\}, \tag{1.1}$$

where C is a copula function of $Y_1, \ldots, Y_d$, linking together the marginal distributions to form the joint distribution. When all the variables are ordinal, then from Sklar (1959), there is a bivariate copula such that

$$P(Y_1 \leq y_1, \ldots, Y_d \leq y_d) = C(F_1(y_1), \ldots, F_d(y_d)),$$

where F_i is the cdf of Y_i with jumps at $0, 1, \ldots, K-1$. When all variables are continuous, C is unique and fully captures the dependence in $\mathbf{Y}$. In that case, the d-dimensional joint density f of $\mathbf{Y}$ can be written as

$$f(y_1, ..., y_d) = c\left\{F_1(y_1), ..., F_d(y_d)\right\} \times \prod_{i=1}^{d} f_i(y_i), \tag{1.2}$$

where $c(u_1, \ldots, u_d)$ is the copula density obtained by taking the partial derivatives of C with respect to each argument. When all (or some) of the variables are ordinal, Equation (1.2) holds but the copula C is not unique and the cdf F_i of Y_i has jumps at $0, 1, \ldots K-1$, for $i = 1, \ldots, d$.

From Equations (1.1) and (1.2), one can see that the joint distribution can be split into parts containing marginal information and a part containing the dependence characteristics. Hence, one can model the marginal distributions and the dependence structure separately. This provides a greater modelling flexibility compared to standard multivariate distributions. Given a copula, one can also construct many new multivariate distributions by selecting dif-

ferent distributions for the margins.

A large number of parametric copula families has been proposed, with the most commonly used families being elliptical copulas (e.g., Gaussian and Student t-copulas) and Archimedean copulas (e.g., Gumbel, Clayton and Frank copulas). Below we describe the parametric copulas used in this thesis, along with the conversions between their parameters and the scale-free dependence measures Kendall's tau (τ) and Spearman's rho (ρ_S). Figure 1.3 displays the contour plots of these copulas with standard normal margins when $\tau = 0.2, 0.5,$ and 0.8.

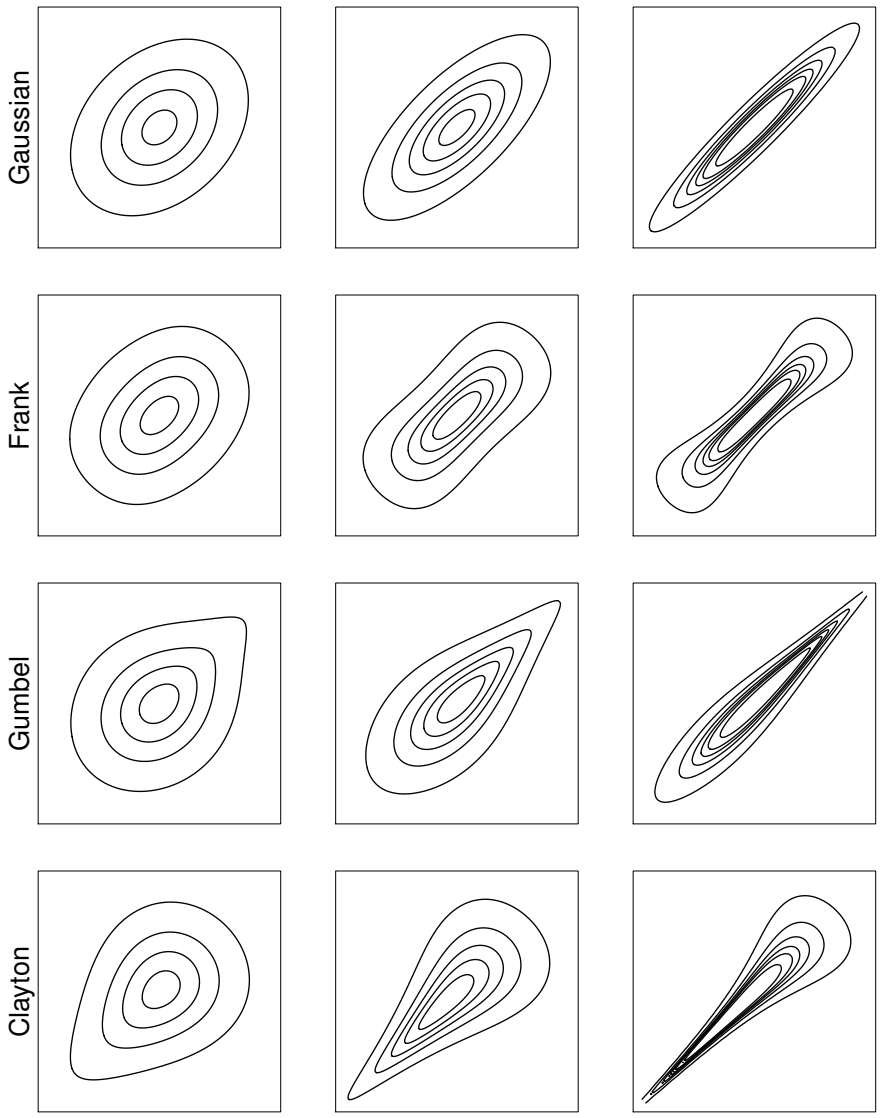
- The **Gaussian** copula is derived from the multivariate normal distribution. Its parameter is determined by the correlation coefficient ρ.

$$\begin{aligned} C_\rho(u,v) &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-(s^2-2\rho st+t^2)}{2(1-\rho^2)}\right\} dt\, ds \\ &= \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), \quad -1 \le \rho \le 1 \end{aligned}$$

The parameter conversions for the Gaussian copula are given by

$$\tau = \frac{2}{\pi} \arcsin(\rho) \quad \text{and} \quad \rho_s = \frac{6}{\pi} \arcsin(\frac{\rho}{2}).$$

Figure 1.3: Contour plots Gaussian, Clayton, Frank and Gumbel copulas with standard normal margins when $\tau = 0.2$ (left panel), 0.5 (middle panel) and 0.8 (right panel).



- The **Frank** copula has no tail dependence, and is given by

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\{1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\}, \qquad \theta \in \mathbb{R}\backslash\{0\}.$$

  The conversions to Kendall's tau and Spearman's rho are given by

$$\tau = 1+\frac{4}{\theta}\left\{D_1(\theta)-1\right\} \quad \text{and} \quad \rho_s = 1-\frac{12}{\theta}\left\{D_1(\theta)-D_2(\theta)\right\},$$

  where $D_j(\theta) = \dfrac{j}{\theta^j}\int_0^\theta \dfrac{t^j}{e^t-1}\,dt$ is the Debye function.

- The **Gumbel** copula exhibits upper tail dependence, and is given by

$$C_\theta(u,v) = exp[-\{(-\ln u)^\theta + (-\ln v)^\theta\}^{\frac{1}{\theta}}], \qquad \theta \in [1,\infty).$$

  The Kendall's tau for the Gumbel copula is given by

$$\tau = 1-\frac{1}{\theta}.$$

  There is no closed form solution for Spearman's rho.

- The **Clayton** copula exhibits lower tail dependence, and is given by

$$C_\theta(u,v) = (u^{-\theta}+v^{-\theta}-1)^{-\frac{1}{\theta}}, \qquad \theta > 0.$$

The dependence parameter of the Clayton copula can be converted to Kendall's tau using

$$\tau = \frac{\theta}{\theta + 2}.$$

The form of Spearman's rho is complicated.

Further reading on copulas and their properties can be found in Nelsen (2006) and Joe (2014).

1.3 Thesis Outline

The structure of this thesis is as follows. Chapter 2 describes the model for factor copula scores regression and outlines data generation. Chapter 3 reviews the dependence measures for multivariate ordinal data and investigates the robustness of these measures to distributional assumptions. Chapter 4 presents estimation of factor scores under the traditional factor model and factor copula model. This chapter also provides a detailed study of factor scores regression under different marginal and joint distributions. Chapter 5 summarizes our main findings and outlines future work.

Chapter 2

Factor Copula Scores Regression: Model and Simulation

This chapter briefly reviews factor model and factor scores regression for multivariate ordinal data under the multivariate normality assumption for the latent variables, and presents the factor copula scores regression model which allows more general multivariate distributions to describe the dependence relationships among the latent and manifest variables. After introducing the notation in Section 2.1, we describe the factor scores regression in Section 2.2 under the multivariate normality assumption. The ways to incorporate more general distributions are discussed in Section 2.3 using factor copulas. In Section 2.4,

we outline a simulation algorithm for data generation under these models.

2.1 Notation

Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ denote a d-dimensional random vector of ordinal variables, which arises from a discretization of $\mathbf{Y}^* = (Y_1^*, \ldots, Y_d^*)$, the vector of underlying continuous unobserved random variables. Let F and F^* represent the d-dimensional cumulative distribution functions of $\mathbf{Y}$ and $\mathbf{Y}^*$, respectively. Further, denote by $U_i = F_i(Y_i)$ and $U_i^* = F_i^*(Y_i^*)$ the marginal cumulative distribution functions of Y_i and Y_i^*, for $i = 1, \ldots, d$. Suppose the dependence among the ordinal variables $Y_1, \ldots, Y_d$, or equivalently of $Y_1^*, \ldots, Y_d^*$ is driven by p unobserved common factors $\mathbf{Z} = (Z_1, \ldots, Z_p)$. We assume that $Z_1, \ldots, Z_p$ are independent and are associated with a univariate covariate X. Throughout the thesis, the statistical realizations of these random variables are denoted by the corresponding lower case letters.

2.2 Traditional Factor Model

For a d-dimensional multivariate normal vector $\mathbf{Y}^*$ and a p-dimensional vector $\mathbf{Z}$, the traditional factor model is given by

$$\mathbf{Y}^* = \mathbf{L}\ \mathbf{Z} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are the vectors of common and specific factors, respectively. The vector $\mathbf{Y}^*$ is linear with respect to the common and specific factors, and the dependence among its components is accounted by $\mathbf{Z}$. For any randomly selected individual, it is assumed that the common factors and the specific factors are independent of one another. It is further assumed that the components of $\mathbf{Z}$ are independent and normally distributed with mean zero and unit variance, i.e,

$$\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p),$$

where $\mathbf{I}_p$ is a p-dimensional identity matrix. Similarly, the components of $\boldsymbol{\epsilon}$ are assumed to have a normal distribution with mean zero and finite variance, yielding

$$\boldsymbol{\epsilon} \sim N_d(\mathbf{0}, \boldsymbol{\Psi}),$$

where $\boldsymbol{\Psi}$ is a diagonal matrix of residual (specific factor) variances.

Factor analysis concerns the covariance structure of $\mathbf{Y}^*$, which can be decomposed as

$$\begin{aligned} \text{Cov}(\mathbf{Y}^*) = \Sigma &= \text{Cov}(\mathbf{L}\ \mathbf{Z} + \boldsymbol{\epsilon}) \\ &= \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}. \end{aligned}$$

The loading matrix $\mathbf{L}$ describes the relationship between $\mathbf{Y}^*$ and $\mathbf{Z}$. Specifically, the $(i, j)^{\text{th}}$ entry of $\mathbf{L}$ represents the covariance between the i^{th} variable

and the j^{th} factor, i.e., $\ell_{ij} = \text{Cov}(Y_i^*, Z_j)$.

One problem that arises in the model in Equation (2.1) is the indeterminacy of the factor solutions. For any given analysis, there can be many estimates of the factor scores that satisfy the specified model. The common factors and the loading matrix are unique up to orthogonal transformations. To see this, let $\mathbf{P}$ denote an orthogonal matrix. Since $\mathbf{P}\,\mathbf{P}' = \mathbf{P}'\,\mathbf{P} = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix, then, $\mathbf{LZ} = \mathbf{L}\,\mathbf{P}\,\mathbf{P}'\,\mathbf{Z} = \mathbf{L}^\star\,\mathbf{Z}^\star$. The identifiability of factors is a challenging problem because an individual who receives a high score on one set of factor scores and could have a low score on another set of factor scores. This makes interpretation of the individual scores very difficult. Typically, one considers various rotations (see Chapter 4) to reach a set of factors that are easy to interpret for a given problem.

2.3 Factor Scores Regression

Regression models involving latent variables have been widely studied; see, for instance Moustaki and Knott (2000); Moustaki (2003); Ines Devlieger (2016). Here, we consider a special type of latent variable regression model where the response is latent and the covariate is observed. Since the latent response is assumed to arise from a factor model, this regression model is referred to as *factor scores regression*.

Suppose each common factor Z_j in Equation (2.1) is related to some ob-

served covariate X through a simple linear model

$$Z_j = \beta_{0j} + \beta_{1j}\ X + \varepsilon_j, \tag{2.2}$$

where ε_j is the random error having $N(0, \sigma^2)$ distribution, and (β_{0j}, β_{1j}) is the parameter vector describing intercept and slope of the linear relation. In genetic association studies, X typically represents the genotype information at a particular marker.

Substituting (2.2) in (2.1) gives an extended form of the traditional factor analysis model that includes the covariate X,

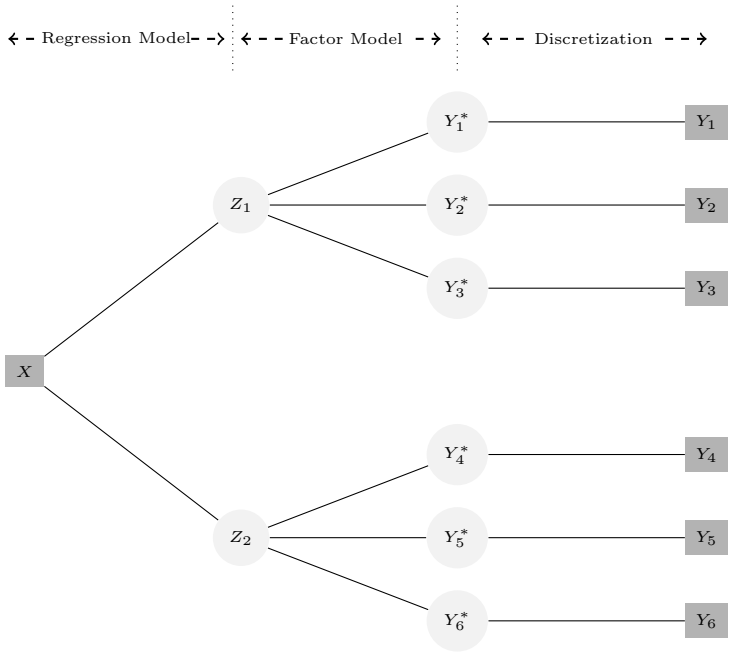
$$\mathbf{Y}^* = \mathbf{\Lambda} + \mathbf{B}^*\ X + \boldsymbol{\zeta}, \tag{2.3}$$

where $\mathbf{\Lambda} = \mathbf{L}\ \boldsymbol{\beta_0}$ and $\mathbf{B}^* = \mathbf{L}\ \boldsymbol{\beta_1}$, with $\boldsymbol{\beta_0} = (\beta_{01}, \ldots, \beta_{0p})'$ and $\boldsymbol{\beta_1} = (\beta_{11}, \ldots, \beta_{1p})'$ and the error term is represented by $\boldsymbol{\zeta}$.

For a randomly selected individual, we assume that X is related to $\mathbf{Y}^*$ indirectly through $\mathbf{Z}$, and that the common factors $\mathbf{Z}$ account for all the dependencies among $\mathbf{Y}^*$, so that if $\mathbf{Z}$ is held fixed, the observed variables are conditionally independent. Figure (2.1) shows the relationships that in an example of six ordinal variables, two factors and a single covariate. Note that X is linked to the manifest variables indirectly through correlations among the $\mathbf{Y}^*$ variables accounted for by Z_1 and Z_2.

Considering only the factor model, the joint distribution of the (continuous)

Figure 2.1: A graphical illustration of variable relationships in the case of six manifest variables, two factors and one covariate.



manifest variables can be written as

$$\begin{aligned} f(\mathbf{y}^*) &= \int_{\mathbb{R}^p} \prod_{i=1}^{d} f(y_i^* \mid z_1, \ldots, z_p) dF_{\mathbf{Z}} \\ &= \int_{\mathbb{R}^p} \prod_{i=1}^{d} f(y_i^* \mid z_1, \ldots, z_p) dF_{Z_1} \ldots dF_{Z_p}. \end{aligned} \tag{2.4}$$

This corresponds to the orthogonal factor model where common factors are assumed to be independent of each other.

Furthermore, the conditional density of the manifest variables given a co-

variate value $X = x$ can be separated into parts involving the factor model in Equation (2.1) and the regression model in Equation (2.2), i.e.,

$$\begin{aligned} f(\mathbf{y}^* \mid x) &= \int_{\mathbb{R}^p} f(\mathbf{y}^*, \mathbf{z} \mid x)\, d\mathbf{z} = \int_{\mathbb{R}^p} f(\mathbf{y}^* \mid \mathbf{z},\ x)\ f(\mathbf{z} \mid x)\, d\mathbf{z} \\ &= \int_{\mathbb{R}^p} f(\mathbf{y}^* \mid \mathbf{z})\ f(\mathbf{z} \mid x)\, d\mathbf{z} \\ &= \int_{\mathbb{R}^p} \left[\prod_{i=1}^{d} f(y_i^* \mid z_1, \ldots, z_p) \right]\ f(z_1 \mid x) \ldots\ f(z_p \mid x)\, d\mathbf{z}. \end{aligned}$$

This factorization plays an important role in obtaining a general model as one can use arbitrary distributions to describe the conditional relationships appearing in the factorization.

2.4 Factor Copula Models

Factor copula models are conditional independence models given latent variables (Krupskii, 2014). They provide an extension of the traditional factor model by allowing arbitrary bivariate linking copulas to define the bivariate relationships among the components of $\mathbf{Y}^*$ and $\mathbf{Z}$. When the linking copulas are all Gaussian, it corresponds to the traditional factor model.

Let $U_i = F_i(Y_i)$, $i = 1, \ldots, d$, be the uniform transformed manifest variables, and $V_j = F(Z_j)$, $j = 1, \ldots, p$, be the common factors in uniform scale.

Then, the general factor copula model is given by

$$C(u_1, \ldots, u_d) = \int_{[0,1]^p} \prod_{i=1}^{d} C_{i|V_1,\ldots,V_p}(u_i|v_1, \ldots, v_p)\ dv_1 \ldots dv_p, \tag{2.5}$$

where $C(u_1, ...u_d)$ is the d-dimensional copula describing the joint distribution of $(U_1, \ldots, U_d)$ and $C_{i|v_1 \ldots v_p}$ is the bivariate copula linking the observed variables U_i to the factors $v_1, \ldots, v_p$. Using this representation, different types of dependence and tail asymmetry can be modelled by appropriate choices of bivariate linking copulas. There are no constraints in the choices of copula families. However, some copula families can lead to more dependence in the joint tail than with a Gaussian copula.

Here, we review the one- and two-factor copula models for continuous manifest variables, and then discuss their extension to ordinal manifest variables.

When $p = 1$, Equation (2.5) yields the one-factor copula model for continuous variables, i.e.,

$$C(u_1, \ldots, u_d) = \int_0^1 \prod_{i=1}^{d} C_{i|V_1}(u_i \mid v_1)\ d_{v_1}, \tag{2.6}$$

where

$$C_{i|v_1}(u_i \mid v_1) = \frac{\partial}{\partial v_1} C_{V_1 i}(u_i, v_1). \tag{2.7}$$

When $p = 2$, Equation (2.5) becomes;

$$\begin{aligned} C(u_1, \ldots, u_d) &= \int_0^1 \int_0^1 \prod_{i=1}^{d} C_{i|V_1,V_2}(u_i \mid v_1, v_2)\, d_{v_1} d_{v_2}, \\ &= \int_0^1 \int_0^1 \prod_{i=1}^{d} C_{i|V_2,V_1}(C_{i|V_1}(u_i \mid v_1) \mid v_2)\, d_{v_1} d_{v_2}, \end{aligned} \tag{2.8}$$

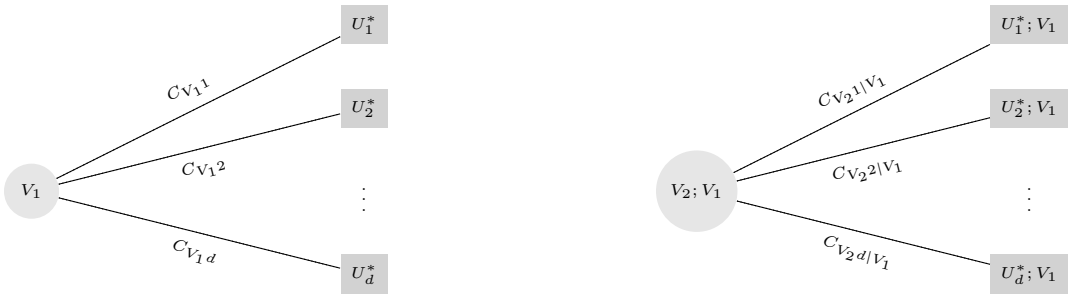
which is the two-factor copula model for continuous variables. When V_1 and V_2 are independent (orthogonal), the conditional distribution $C_{i|V_1,V_2}$ can be written as

$$\begin{aligned} C_{i|V_1,V_2}(u \mid v_1, v_2) &= P(U_i \leq u \mid V_1 = v_1, V_2 = v_2) \\ &= \frac{\partial}{\partial v_2} P(U_i \leq u, V_2 \leq v_2 \mid V_1 = v_1). \\ &= \frac{\partial}{\partial v_2} C_{V_2 i;V_1}(v_2, C_{i|V_1}(u \mid v_1)) \\ &= C_{i|V_2;V_1}(C_{i|V_1}(u \mid v_1) \mid v_2). \end{aligned}$$

This model is illustrated in Figure 2.2, which provides guidance in data generation (see Algorithm 2).

For ordinal manifest variables, the distribution function F_i of Y_i is a step function with jumps at $0, 1, \ldots, K-1$, where the latter represent the category labels. The dependence between each observed variable and the latent variable

Figure 2.2: A graphical representation of the two-factor copula model with d uniform variables. The circles represent latent variables and the boxes represent manifest variables in uniform scale.



can be written in terms of copulas as

$$P(V_j \leq v, Y_i \leq y) = C_{\mathbf{V}_j i}(V_j, F_i(y)), \quad 0 < v < 1.$$

It follows that the joint probability mass function (pmf) is given as

$$\begin{aligned}\pi_d(\mathbf{y}) &= P(Y_1 = y_1, \ldots, Y_d = y_d) \\ &= \int_{[0,1]^p} \prod_{i=1}^{d} P(Y_i = y_i \mid V_1 = v_1, \ldots, V_p = v_p)\ dv_1, \ldots, dv_p.\end{aligned}$$

If there is one latent variable explaining the dependence among $\mathbf{Y}$, then the one-factor copula model is expressed in terms of the joint probability mass

function $\pi_{1:d}(\mathbf{y})$ as

$$\begin{aligned}\pi_{1:d}(\mathbf{y}) &= \int_0^1 \prod_{i=1}^{d} P(Y_i = y_i \mid V_1 = v)\, dv \\ &= \int_0^1 \prod_{i=1}^{d} \big[C_{i|V_1}(F_i(y_i) \mid v_1) - C_{i|V_1}(F_i(y_i - 1) \mid v_1)\big] dv_1.\end{aligned}$$

Similarly, the two-factor copula model for ordinal data is given as

$$\begin{aligned}\pi_{1:d}(\mathbf{y}) &= \int_0^1 \int_0^1 \prod_{i=1}^{d} P(Y_i = y_i \mid V_1 = v_1, V_2 = v_2)\, dv_1 dv_2. \\ &= \int_0^1 \int_0^1 \prod_{i=1}^{d} \Big[C_{i|V_2;V_1}(F_{i|V_1}(y_i \mid v_1) \mid v_2) \\ &\qquad - C_{i|V_2;V_1}(F_{i|V_1}(y_i - 1 \mid v_1) \mid v_2)\Big] dv_1 dv_2.\end{aligned}$$

2.5 Data Generation

In line with our objectives, we consider a data generating model that accounts for both the regression model and the factor model for multivariate ordinal data, in the respective order.

As a first step, we simulate covariate information for n independent subjects. To mimic the setting of genetic association studies, we consider X to be

categorical with three levels $X \in \{0, 1, 2\}$, which represents the genotype at a particular genetic marker. Thus, we cluster n subjects based on their genotype group. For each $j = 1, \ldots, p$, the random error ε_j is generated from a normal distribution with a mean of 0 and standard deviation of σ_j. The true factors Z_j's are then simulated independently using the model in (2.2).

The regression coefficient β_{1j} is related to the Pearson correlation ρ_j of Z_j and X through $\rho_j \times \sigma_j/\sigma_X$, where σ_j and σ_X are their respective standard deviations. In our implementations, we fix the coefficient of determination $Q_j = \rho_j^2$ at a certain level and determine the corresponding slope parameter β_{1j} for $j = 1, \ldots, p$.

Note that for each factor we have $Z_j \mid X = x \sim N(\beta_{0j} + \beta_{1j}x, \sigma_j^2)$. In order for the common factor to have an unconditional standard normal distribution, the mean is computed as

$$\mathrm{E}(Z_j) = \mathrm{E}(\mathrm{E}(Z_j|X)) = \mathrm{E}(\beta_{0j} + \beta_{1j}X) = 0,$$

which implies that

$$\beta_{0j} = -\beta_{1j}\ \mathrm{E}(X).$$

Similarly, the unconditional variance

$$\mathrm{Var}(Z) = \mathrm{Var}(\mathrm{E}(Z|X)) + \mathrm{E}(\mathrm{Var}(Z|X)) = \mathrm{Var}(\mathrm{E}(Z|X)) + \sigma_j^2 = 1,$$

yields

$$\sigma_j^2 = 1 - \text{Var}(\beta_{0j} + \beta_{1j}X) = 1 - \beta_{1j}^2 Var(X).$$

In order to generate the manifest variables, we need to define a data generating model that share the p common factors given in (2.2). That is, we need to "link" the regression model and the factor model to obtain manifest variables.

For this we first convert each Z_j to uniform scale using the corresponding probability integral transformation, i.e., the unconditional distribution.

$$V_j = P(Z_j \leq z) = \sum_{x=0}^{2} P(X = x)\ P(Z_j \leq z \mid X = x),$$

for $j = 1, \ldots, p$. Given V_j, we specify a sequence of parametric bivariate linking copulas (i.e., copula families and parameters) to generate uniform data $\mathbf{U}$. In this step, the relation in Equation (2.7) is commonly used. We then apply the inverse of the probability integral transformation to convert each U_i to Y_i^* using a continuous distribution. If the manifest variables are continuous then the resulting $\mathbf{Y}^*$ gives the desired data matrix. For ordinal manifest variables, we use the marginal probabilty mass functions (pmf's) to discretize $\mathbf{Y}^*$ and get $\mathbf{Y}$.

In our implementations, we focus our attention to two-factor copula models with orthogonal factors and only one type of copula family in the bivariate

linking copulas. However, one can also define settings with dependent factors or where linking copulas belong to different families.

The specification of the factor copula model requires setting a dependence parameter for each bivariate linking copula. For two or more common factors, the copula parameters represent partial correlations for all factors but the first one. Although the interpretation of the copula parameters does not coincide with that of the entries of the loading matrix $\mathbf{L}$, they can be converted to one another. Since our interest is in robustness of the traditional factor analysis, and since $\mathbf{L}$ has a more direct interpretation, we fix $\mathbf{L}$ and determine the copula parameters accordingly when generating data.

In order for the different copula families to have the same strength of dependence, measured in terms of Kendall's tau, the correlations (loading matrix entries) are first converted to partial correlations using the relations

$$\theta_{i1}^{(G)} = \ell_{i1} \quad \text{and} \quad \theta_{i2}^{(G)} = \frac{\ell_{i2}}{\sqrt{1-\ell_{i1}^2}},$$

where ℓ's are the loadings and $\theta^{(G)}$'s are the copula parameters under Gaussian bivariate linking copulas. We convert these parameters to Kendall's tau and then from Kendall's tau to the corresponding copula parameters θ's using the one-to-one transformations in Chaptert 1.

The uniform variables generated from the two-factor copula model are then discretized into four ordered categories (denoted as $0, 1, 2, 3$ to reflect a sever-

ity score from mild to severe) by fixing the cut-off values for each marginal distribution based on their respective pmf's. Different marginal distributions are considered to reflect varying levels of skewness. The resulting multivariate ordinal data $\mathbf{Y} = (Y_1, \ldots, Y_d)$ for n subjects represent the ordinal responses for a set of d questions in a diagnostic questionnaire. The observed data on each subject hence consists of $(X, Y_1, \ldots, Y_d)$ for a random sample of size n.

The algorithms used to simulate ordinal data from a one-factor copula model and two-factor copula model are described in Algorithm 1 and Algorithm 2, respectively. Further details can be found in Joe (2014).

Algorithm 1 Simulating One-Factor Copula Model for Ordinal Data

1. Given X, generate ε and obtain Z
2. $V \leftarrow F_Z(Z)$
3. Generate $w_1, ..., w_d$ to be independent U(0,1) random variables
4. $U_i^* \leftarrow C_{i|V}^{-1}(w_i|v)$, for $i = 1, ..., d$
5. $\mathbf{Y}^* \leftarrow F_{U^*}^{-1}(u^*)$
6. Use marginal pmf's to discretize $Y_1^*, \ldots, Y_d^*$.
7. Return multivariate ordinal data $(Y_1, \ldots, Y_d)$

Algorithm 2 Simulating Two-Factor Copula Model for Ordinal Data

1. Given X, generate ε_j and obtain Z_p for $p = 1, 2$
2. $V_k \leftarrow F_{Z_p}(Z_p)$
3. Generate $w_1, ..., w_d$ to be independent U(0,1) random variables.
4. for $i = 1, ..., d$: do
5. $q_{2i} \leftarrow C^{-1}_{i|V_2;V_1}(w_i|v_2)$,
6. $U^*_i \leftarrow C^{-1}_{i|V_1}(q_{2i}|v_1)$,
7. end for
8. $\mathbf{Y}^* \leftarrow F^{-1}_{U^*}(u^*)$
9. Use marginal pmf's to discretize $Y^*_1, \ldots, Y^*_d$
10. Return multivariate ordinal data $(Y_1, \ldots, Y_d)$

Chapter 3

Robustness of Dependence Measures for Ordinal Data

This chapter outlines commonly used association measures for ordinal variables and presents results from a simulation study that compares the discrepancy in the estimates of these dependence measures under different distributional settings.

3.1 Dependence Measures for Ordinal Data

In many applications, the associations between variables are assessed by measuring the strength of dependence. In the case of continuous variables, common measures of dependence are Pearson's correlation, Spearman's rank correlation

and Kendall's tau. These measures, however are not suitable when variables are discrete or measured on the ordinal scale.

In the following, we focus on the Pearson- and Spearman-type dependence measures, since these measures admit a matrix form, and are more natural for factor analysis than a matrix formed by Kendall's tau. We first describe each for continuous variables and then discuss their adaptations to ordinal variables.

3.1.1 Pearson Correlation

The Pearson product-moment correlation coefficient is a measure of the strength of linear dependence between two continuous variables. It takes values between +1 (perfect positive linear correlation) and -1 (perfect negative linear correlation) inclusive. A value of 0 indicates no linear correlation. It is a convenient measure of dependence and is widely used because of its ease of calculation and interpretation. It is measured as the covariance of two variables divided by the product of their standard deviations. Mathematically for any two random variables X and Y, the Pearson correlation coefficient is defined as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{3.1}$$

The sample correlation coefficient, r, is computed as the sample covariance divided by the product of the sample standard deviations, i.e.,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.2}$$

For ordinal variables, one approach for measuring dependence is to assume a meaningful metric so that distance comparisons between values are suitable. Although, the Pearson correlation can be calculated if one treats the ordinal variables as continuous or having metric properties (Bollen and H.Barb, 1981), this approach can produce misleading results especially when used in factor analysis (Francisco P.Holgado-Tello et al., 2008).

Dating back to Pearson (1900), a popular measure for ordinal variables is the *polychoric* correlation. The basic concept of the polychoric correlation coefficient lies in the assumption that two ordinal variables are a discretization of continuous (latent) variables, which jointly have a bivariate normal distribution.

If we denote the rectangles derived from the discretization of the joint normal distribution by A_s (see Figure 1.2), and the bivariate normal density with parameter ρ by ϕ, then the joint probabilities of the discretized joint normal distribution is equivalent to the joint probabilities of the ordinal variables

(Joakim, 2008)

$$\int_{A_s} \phi_\rho(\mathbf{y}^*) d\mathbf{y}^* = P(\mathbf{Y} \in A_s) = P((Y_1, Y_2) \in A_s) \tag{3.3}$$

These rectangles corresponds to the cells or categories of the bivariate ordinal data. The solution to equation (3.3) gives the polychoric correlation coefficient between the Y_1 and Y_2.

The Pearson-type correlation measures have some well-recognized drawbacks. They measure only the strength of linear dependence and can therefore be misleading when dependencies are nonlinear. Furthermore, they are invariant only under linear transformations and can be sensitive to outliers. Therefore, more robust dependence measures are often preferred in practice.

3.1.2 Spearman Rank Correlation

One robust dependence measure is Spearman's rank correlation, which is a nonparametric version of Pearson's correlation calculated on rankings of two variables. The values of Spearman's rho fall between -1 and $+1$, and quantify the strength of the association. Its sign determines the the direction of the association between the two random variables. For any two independent random vectors (X_1, Y_1) and (X_2, Y_2) having an identical distribution with that

of (X, Y), the Spearman rho is defined as,

$$\rho_S = 3(P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]). \quad (3.4)$$

For continuous data, the sample version of Spearman's correlation is expressed as,

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad (3.5)$$

where $d_i = \text{Rank}(X_i) - \text{Rank}(Y_i)$, is the difference between the two ranks of each observation.

The extensions of Spearman's correlation for discrete data have been considered in Neslehova (2004), Denuit and Lambert (2005) and Luo (2011). A discrete version of Spearman's correlation (Luo, 2011) that takes into account the ordinal nature of the random variables is defined as,

$$\tilde{\rho}_S = \frac{\rho_S}{\sqrt{(1 - \sum_{i=1}^{m} p_i^3)(1 - \sum_{j=1}^{n} q_j^3)}}, \quad (3.6)$$

where p_i and q_j are the marginal probabilities of X and Y), respectively, for $i = 1, 2, \ldots m$ and $j = 1, 2, \ldots n$. Furthermore, using the relationship between Pearson's correlation and Spearman's correlation, Luo (2011) defines

$$\tilde{\rho}_{NP} = \frac{6}{\pi} \arcsin(\frac{\rho}{2}), \quad (3.7)$$

as a non-parametric estimate of the polychoric correlation.

In the next section, we investigate the robustness of these measures in a simulation study.

3.2 Numerical Assessments

We conduct a simulation study to compare the discrepancy of the Pearson- and Spearman-type dependence measures for bivariate ordinal data. Our aim is to investigate the impact of copula family and marginal distributions on the estimation of these dependence measures.

3.2.1 Simulation Setting

Since our investigations on polychoric correlation are mainly for factor analysis and factor scores regression, we consider a two-factor copula model in our data generation and use the same datasets in the numerical assessments in Chapter 4.

We consider four copula families to specify the bivariate linking copulas for a two-factor copula model. The Gaussian and Frank copulas represent the settings with symmetrical dependence (i.e., no tail dependence) whiles the Clayton and Gumbel copulas represent scenarios with asymmetrical dependence exhibiting lower and upper tail dependence, respectively. We further

consider three different marginal distributions for the margins to reflect varying levels of skewness. These are listed in Table 3.1. While $p_1(x)$ illustrates the symmetric case, $p_2(x)$ and $p_3(x)$ reflect the left- and right-skewed cases.

Table 3.1: Marginal probability mass functions.

x	0	1	2	3
$p_1(x)$	0.25	0.25	0.25	0.25
$p_2(x)$	0.05	0.15	0.20	0.60
$p_3(x)$	0.60	0.20	0.15	0.05

In all settings, we use the following loading matrix to determine the copula parameters in data generation.

$$\mathbf{L}' = \begin{pmatrix} 0.90 & 0.80 & 0.50 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.90 & 0.80 & 0.50 \end{pmatrix}$$

Given the loading matrix, the data generation is performed following the steps in Section 2.5. Under each setting, we first generate covariate data on X for a single SNP with minor allele frequency (MAF) 10% from the multinomial distribution with probabilites $(0.81, 0.18, 0.01)$. The three cases considered for $n = 200, 500$ and 1000 resulted in the following empirical counts $(167, 30, 3)$, $(410, 78, 12)$ and $(816, 168, 16)$, respectively.

Then, using Algorithm 2, we obtain uniform data $\mathbf{U}$ and transform these to the normal scale via $Y_i^* = \Phi^{-1}(U_i)$, $i = 1, \ldots, d$, where Φ stands for the

standard normal distribution. The data on the continuous normal (latent) variables $\mathbf{Y}^*$ are then discretized into four categories ($0, 1, 2$ and 3) by fixing the cut-off values for each given marginal distribution.

3.2.2 Simulation Results

Under each setting of a copula family and marginal distribution, we run 1000 experiments of sample sizes $n = 200$, 500 and 1000, and obtain estimates of the association measures for both the continuous (latent) data and the ordinal data.

Specifically, we compare the sample estimates of the Pearson correlation (ρ^*) for the continuous (unobserved) variables, the Pearson correlation (ρ) which ignores the data were ordinal, the polychoric correlation ($\tilde{\rho}$) that is numerically calculated from the discretization of a bivariate normal distribution, the Spearman correlation (ρ_S^*) for continuous variables, the Spearman correlation (ρ_S) that ignores the data were ordinal, the Spearman correlation ($\tilde{\rho}_S$) for ordinal variables as defined in Equation (3.6), and the non-parametric correlation ($\tilde{\rho}_{NP}$) given by Equation (3.7).

The estimates are obtained for each pair of variables. Here, we only report the results on the dependence between Y_1 and Y_2. Note that the underlying correlation between the continuous latent variables Y_1^* and Y_2^* is given by $\ell_{11}\,\ell_{21}+\ell_{12}\,\ell_{22} = 0.90\,(0.80)+0.01\,(0.01) = 0.7201$, and is used as a benchmark in our evaluations.

The estimation results under the Gaussian, Frank, Gumbel and Clayton copulas are summarized in Tables 3.2, 3.3, 3.4 and 3.5, respectively. For continuous latent variables, the estimates of Pearson's correlation are very close to the true value, especially when the underlying dependence is Gaussian. This conclusion holds even when the margins have moderate skewness. When the underlying dependence is non-Gaussian (e.g., Clayton, Gumbel and Frank), the estimates of ρ^* are only slightly lower than the true value.

The skewness in the margins has a negligible effect on the polychoric correlation estimates under the Gaussian and Frank copulas, but can lead to overestimation or underestimation under the Clayton and Gumbel copulas. On the other hand, when the marginal distributions are symmetric, the impact of asymmetry in the copula seems negligible.

In order to evaluate the significance of the observed overestimation or underestimation, we constructed confidence intervals using the average correlation estimates and checked if the true value is contained in the intervals. Due to the restricted range of the correlation parameter, we use the logit scale in our assessments. That is, for each run we first obtained the logit transformed estimates using $w = log(r/(1-r))$ and calculated the mean and standard error. We then constructed 95% confidence intervals using mean $\pm$ 2 $\times$ standard error. The cases where the confidence intervals do not contain the true value in logit scale are reported in bold, suggesting a significant violation of robustness.

Based on the confidence intervals, the estimates of Pearson's and Spearman's correlation ignoring the data were ordinal consistently violate robustness under all copula families and marginal distributions. The polychoric correlation and discrete versions of Spearman's rho yield more accurate estimates. Overall, the true value in logit scale is contained in the confidence intervals except under the Gumbel and Clayton copula families with especially left-skewed marginal distributions.

Table 3.2: Mean and standard error (in parenthesis) of the Pearson- and Spearman-type correlation estimates for `Item 1` and `Item 2` under the Gaussian copula model with different marginal distributions and sample size, over 1000 Monte-Carlo samples.

		Pearson Correlations			Spearman Correlations			
Margins	n	ρ^*	ρ	$\tilde{\rho}$	ρ_S^*	ρ_S	$\tilde{\rho}_S$	$\tilde{\rho}_{NP}$
Symmetric	200	0.718 (0.036)	0.643 (0.047)	0.716 (0.046)	0.699 (0.041)	0.643 (0.047)	0.686 (0.050)	0.703 (0.049)
	500	0.720 (0.022)	**0.646 (0.028)**	0.719 (0.028)	0.703 (0.024)	**0.646 (0.028)**	0.689 (0.030)	0.706 (0.029)
	1000	0.720 (0.015)	**0.647 (0.020)**	0.720 (0.020)	0.703 (0.017)	**0.647 (0.020)**	0.691 (0.021)	0.707 (0.021)
L. Skewed	200	0.718 (0.036)	**0.606 (0.057)**	0.712 (0.055)	0.699 (0.041)	**0.577 (0.058)**	0.747 (0.075)	0.762 (0.073)
	500	0.720 (0.022)	**0.610 (0.034)**	0.718 (0.032)	0.703 (0.024)	**0.583 (0.034)**	0.754 (0.044)	0.769 (0.043)
	1000	0.720 (0.015)	**0.612 (0.024)**	0.719 (0.023)	0.703 (0.017)	**0.583 (0.025)**	0.755 (0.032)	0.770 (0.031)
R. Skewed	200	0.718 (0.036)	**0.609 (0.056)**	0.716 (0.054)	0.699 (0.041)	**0.581 (0.057)**	0.752 (0.074)	0.767 (0.072)
	500	0.720 (0.022)	**0.612 (0.034)**	0.720 (0.032)	0.703 (0.024)	**0.584 (0.035)**	0.756 (0.046)	0.771 (0.044)
	1000	0.720 (0.015)	**0.611 (0.024)**	0.719 (0.023)	0.703 (0.017)	**0.582 (0.024)**	0.753 (0.032)	0.769 (0.030)

Table 3.3: Mean and standard error (in parenthesis) of the Pearson- and Spearman-type correlation estimates for `Item 1` and `Item 2` under the Frank copula model with different marginal distributions and sample size, over 1000 Monte-Carlo samples.

		Pearson Correlations			Spearman Correlations			
Margins	n	ρ^*	ρ	$\tilde{\rho}$	ρ_S^*	ρ_S	$\tilde{\rho}_S$	$\tilde{\rho}_{NP}$
Symmetric	200	0.667 (0.038)	0.671 (0.043)	0.740 (0.043)	0.707 (0.038)	0.671 (0.043)	0.716 (0.046)	0.732 (0.045)
	500	**0.668 (0.023)**	0.674 (0.025)	0.742 (0.026)	0.710 (0.022)	0.674 (0.025)	0.718 (0.027)	0.735 (0.026)
	1000	0.667 (0.016)	**0.675 (0.018)**	0.744 (0.018)	0.711 (0.016)	**0.675 (0.018)**	0.720 (0.019)	0.736 (0.018)
L. Skewed	200	0.667 (0.038)	**0.577 (0.057)**	0.688 (0.053)	0.707 (0.038)	**0.599 (0.057)**	0.775 (0.073)	0.789 (0.071)
	500	**0.668 (0.023)**	0.581 (0.035)	0.694 (0.032)	0.710 (0.022)	**0.605 (0.034)**	0.783 (0.044)	0.797 (0.042)
	1000	**0.667 (0.016)**	**0.583 (0.026)**	0.695 (0.023)	0.711 (0.016)	**0.606 (0.024)**	0.784 (0.032)	0.798 (0.030)
R. Skewed	200	0.667 (0.038)	**0.580 (0.057)**	0.692 (0.052)	0.707 (0.038)	**0.602 (0.055)**	0.779 (0.071)	0.793 (0.069)
	500	0.668 (0.023)	**0.584 (0.036)**	0.696 (0.032)	0.710 (0.022)	**0.606 (0.035)**	0.785 (0.045)	0.799 (0.043)
	1000	**0.667 (0.016)**	**0.582 (0.026)**	0.695 (0.023)	0.711 (0.016)	**0.605 (0.024)**	0.783 (0.032)	0.797 (0.031)

Table 3.4: Mean and standard error (in parenthesis) of the Pearson- and Spearman-type correlation estimates for `Item 1` and `Item 2` under the Gumbel copula model with different marginal distributions and sample size, over 1000 Monte-Carlo samples.

		Pearson Correlations			Spearman Correlations			
Margins	n	ρ^*	ρ	$\tilde{\rho}$	ρ_S^*	ρ_S	$\tilde{\rho}_S$	$\tilde{\rho}_{NP}$
Symmetric	200	0.710 (0.040)	0.638 (0.049)	0.711 (0.049)	0.691 (0.044)	0.638 (0.049)	0.680(0.052)	0.697 (0.051)
	500	0.712 (0.024)	**0.640 (0.029)**	0.714 (0.029)	0.695 (0.026)	**0.640 (0.029)**	0.683 (0.032)	0.700 (0.032)
	1000	0.713 (0.017)	**0.641 (0.021)**	0.715 (0.020)	0.695 (0.019)	**0.641 (0.021)**	0.684 (0.022)	0.701 (0.022)
L. Skewed	200	0.710 (0.040)	**0.523 (0.062)**	0.635 (0.061)	0.691 (0.044)	**0.522 (0.061)**	0.676 (0.080)	0.692 (0.078)
	500	0.712 (0.024)	**0.528 (0.039)**	**0.641 (0.037)**	0.695 (0.026)	**0.528 (0.037)**	0.683 (0.047)	0.700 (0.047)
	1000	0.713 (0.017)	**0.528 (0.028)**	**0.642 (0.027)**	0.695 (0.019)	**0.528 (0.026)**	0.684 (0.034)	0.701 (0.033)
R. Skewed	200	0.710 (0.040)	0.685 (0.049)	0.788 (0.046)	0.691 (0.044)	0.627 (0.055)	0.812 (0.071)	0.824 (0.068)
	500	0.712 (0.024)	0.688 (0.031)	0.792 (0.028)	0.695 (0.026)	0.630 (0.034)	0.815 (0.043)	0.828 (0.041)
	1000	0.713 (0.017)	0.688 (0.021)	0.791 (0.020)	0.695 (0.019)	0.629 (0.024)	0.814 (0.031)	0.827 (0.029)

Table 3.5: Mean and standard error (in parenthesis) of the Pearson- and Spearman-type correlation estimates for `Item 1` and `Item 2` under the Clayton copula model with different marginal distributions and sample size, over 1000 Monte-Carlo samples.

		Pearson Correlations			Spearman Correlations			
Margins	n	ρ^*	ρ	$\tilde{\rho}$	ρ_S^*	ρ_S	$\tilde{\rho}_S$	$\tilde{\rho}_{NP}$
Symmetric	200	0.701 (0.041)	0.658 (0.047)	0.729 (0.046)	0.702 (0.043)	0.658 (0.047)	0.702 (0.050)	0.718 (0.049)
	500	0.702 (0.025)	**0.661 (0.029)**	0.733 (0.029)	0.706 (0.026)	**0.661 (0.029)**	0.705 (0.031)	0.722 (0.030)
	1000	0.701 (0.017)	**0.661 (0.020)**	0.733 (0.020)	0.705 (0.018)	**0.661 (0.020)**	0.705 (0.022)	0.722 (0.021)
L. Skewed	200	0.701 (0.041)	**0.767 (0.040)**	0.862 (0.035)	0.702 (0.043)	**0.699 (0.048)**	0.905 (0.062)	0.912 (0.058)
	500	0.702 (0.025)	**0.772 (0.024)**	**0.867 (0.020)**	0.706 (0.026)	**0.704 (0.029)**	0.912 (0.038)	0.919 (0.035)
	1000	0.701 (0.012)	**0.774 (0.017)**	**0.868 (0.015)**	0.705 (0.018)	**0.705 (0.021)**	0.913 (0.027)	0.920 (0.026)
R. Skewed	200	0.701 (0.041)	0.429 (0.067)	0.552 (0.067)	0.702 (0.043)	0.474 (0.066)	0.614 (0.085)	0.631 (0.086)
	500	0.702 (0.025)	0.432 (0.043)	0.556 (0.043)	0.706 (0.026)	0.477 (0.041)	0.618 (0.055)	0.636 (0.053)
	1000	0.701 (0.017)	0.430 (0.031)	0.553 (0.031)	0.705 (0.018)	**0.475 (0.030)**	0.615 (0.039)	0.632 (0.038)

3.3 Summary

For the scenarios considered in our simulations, the polychoric correlation estimates are found to be very close to the corresponding correlation of the underlying latent variables. As expected, the polychoric correlation consistently performed well when the underlying distribution is Gaussian.

We observe that the polychoric correlation estimates are not severely affected by the skewness of the marginal distributions under copulas with symmetric dependence. However, under the left-and right-skewed marginal distributions, the polychoric correlation estimates inflates or deflates slightly when the underlying copula family has an asymmetric dependence structure. Based on our investigations, not reported here, we also observe that the discrepancy between the true and polychoric correlation estimates increases as the degree of skewness in the probability marginal functions. Overall, the impact of marginal distributions is found to be negligible when the underlying copula exhibit symmetric dependence, and the impact of copula is negligible when the underlying marginal distributions are symmetric.

The estimates of Spearman's rho are not directly comparable to those of Pearson's correlation, except for the non-parametric rank correlation ($\tilde{\rho}_{NP}$). The magnitude of the latter was overall higher than the polychoric correlation, indicating that its use in factor analysis may not always yield accurate results. Therefore, we mainly investigate polychoric correlation-based factor analysis in the next chapter.

Chapter 4

Robustness of Factor Analysis and Factor Scores Regression

In this chapter, we briefly review the estimation of the traditional factor model and factor scores regression, and discuss the same for the factor copula model and factor copula scores regression. We then investigate the robustness of these procedures via simulations under settings with asymmetric dependence and skewed marginal distributions.

4.1 Estimation of the Traditional Factor Model

The estimation of the traditional factor model amounts to the estimation of the factor model parameters (loading matrix), which are the coefficients that

gives the weights between the manifest variables and the common factors. For the given number of factors, we estimate the loading matrix via the method of maximum likelihood. This procedure is commonly used to fit a factor model.

A number of different approaches for computing the factor scores have been developed in the factor analysis literature and one must choose from the available options. These approaches include regression, Bartlett's method and the Anderson and Rubin's method among others (Hershberger, 2005).

The regression method uses the maximum likelihood to estimate the factor loadings. This procedure uses the correlations among the manifest variables as well as the correlations between the manifest variables and the common factors to compute the factor scores. If we consider the joint distribution of the data $\mathbf{Y}$ and the factor Z as

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{L}\mathbf{L}' + \mathbf{\Psi} & \mathbf{L} \\ \mathbf{L}' & \mathbf{I} \end{pmatrix} \right]$$

we can calculate conditional expectation of the Z given the manifest variables Y as

$$\mathrm{E}(Z|Y) = \mathbf{L}'(\mathbf{L} \;\; \mathbf{L}' + \mathbf{\Psi})^{-1} \;\; (\mathbf{Y} - \mu).$$

By substituting in the estimates for $\mathbf{L}$ and $\mathbf{\Psi}$, we get the estimator as

$$\hat{\mathbf{Z}} = \hat{\mathbf{L}}'(\hat{\mathbf{L}} \ \ \hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})^{-1} \ \ (\mathbf{Y} - \bar{\mathbf{y}}).$$

In our implementations, we use the default regression option in R psych package (Revelle, 2014).

From the fitted factor model, we estimate the scores $\widehat{\mathbf{Z}} = (\widehat{Z}_1, \ldots, \widehat{Z}_P)$ for n subjects. To improve the interpretability of the factor analysis, we consider no rotation, as well as varimax and oblimin rotations, as commonly used in the psychometrics literature.

4.2 Estimation of the Factor Copula Model

The fitting of factor copula models requires the estimation of the copula parameters for each bivariate linking copula in the model. Due to the latent variables in the model, one typically estimates the model parameters jointly using maximum likelihood. Suppose the copula families in a factor copula model are known and each bivariate linking copula C_{iV_j} is characterized by a univariate copula parameter $\theta_{iV_j} \in \Theta_{iV_j}$. Then, the parameter vector to be (jointly) estimated is $\boldsymbol{\theta} = (\theta_{1V}, \ldots, \theta_{dV})$ for the one-factor copula model, and $\boldsymbol{\theta} = (\theta_{1V_1}, \ldots, \theta_{dV_1}, \theta_{1V_2}, \ldots, \theta_{dV_2})$ for the two-factor copula model.

Consider a random sample $\{Y_{1s}, \ldots, Y_{ds}\}_{s=1}^{n}$ of random variables for n subjects. We first look into the continuous case, and outline the case with known

marginal distributions for notational simplicity. Let $U_{is} = F_i(Y_{is})$, $i = 1, \ldots, d$ denote the probability integral transform of the manifest variables. Then, the log-likelihood function is given by

$$\mathcal{L}(u_1, ...u_d; \boldsymbol{\theta}) = \sum_{s=1}^{n} \log \left\{ c(u_{1s}, \ldots, u_{ds}; \boldsymbol{\theta}) \right\}. \tag{4.1}$$

For a one-factor copula model, the copula density $c(u_{j1}, ...u_{jd}; \boldsymbol{\theta})$ is

$$c(u_{1s}, ...u_{ds}; \boldsymbol{\theta}) = \int_0^1 \prod_{i=1}^{d} c_{iV}(u_{is}, v; \theta_{iV}) dv,$$

which can be approximated using Gauss-Legendre quadrature, i.e.,

$$c(u_{1s}, ...u_{ds}; \boldsymbol{\theta}) \approx \sum_{k=1}^{n_q} w_k \prod_{i=1}^{d} c_{iV}(u_{is}, x_k; \theta_{iV}),$$

where n_q is the number of quadrature points (usually taken as $n = 25$), x_k are the nodes and w_k are the quadrature weights. Similarly, the two-factor

parametric copula density is given by

$$c(u_{1s},\ldots,u_{ds};\boldsymbol{\theta}) = \int_0^1\int_0^1 \prod_{i=1}^{d} c_{iV_2;V_1}(C_{i|V_1}(u_{is}|v_1;\theta_{iV_1}),v_2;\theta_{iV_2}) \times c_{iV_1}(u_{is},v_1;\theta_{iV_1})\, dv_1\, dv_2.$$

An approximation for the two-dimensional integral is given as

$$c(u_{1s},\ldots,u_{ds};\boldsymbol{\theta}) \approx \sum_{k_1=1}^{n_q}\sum_{k_2=1}^{n_q} w_{k_1}w_{k_2}\prod_{i=1}^{d}\{c_{iV_2;V_1}(C_{i|V_1}(u_{is}|x_{k_1};\theta_{iV_1}),x_{k_2};\theta_{iV_2}) \times c_{iV_1}(u_{is},x_{k_1};\theta_{iV_1})\}.$$

In the ordinal case, the major difference is that the density expressions are replaced by differences in cdf's at consecutive points. The log-likelihood function for the multivariate ordinal data is defined in the usual way as

$$\mathcal{L}(y_1,....y_d;\boldsymbol{\theta}) = \sum_{s=1}^{n}\log\left\{\pi_{i:d}(y_{1s},\ldots,y_{ds};\boldsymbol{\theta})\right\}. \tag{4.2}$$

For the one-factor copula model, the pmf of ordinal data is given by

$$\pi_{i:d}(\mathbf{y};\theta) = \int_0^1\prod_{i=1}^{d}\left[C_{i|V_1}(F_i(y_i)\mid v_1,\theta_{iV}) - C_{i|V_1}(F_i(y_i-1)\mid v_1,\theta_{iV})\right]dv_1,$$

which is approximated by

$$\pi_{i:d}(\mathbf{y};\theta) \approx \sum_{k=1}^{n_q} w_k \prod_{i=1}^{d} \left[C_{i|V_1}(F_i(y_i) \mid v_1, \theta_{iV}) - C_{i|V_1}(F_i(y_i-1) \mid v_1, \theta_{iV})\right] dv_1.$$

Similarly, the pmf of the two-factor copula model is defined as

$$\begin{aligned}\pi_{1:d}(\mathbf{y},\theta) &= \int_0^1 \int_0^1 \prod_{i=1}^{d} \Big[C_{i|V_2;V_1}(F_{i|V_1}(y_i \mid v_1;\theta_{iV_1}) \mid v_2;\theta_{iV_2}), \\ &\qquad -C_{i|V_2;V_1}(F_{i|V_1}(y_i-1 \mid v_1;\theta_{iV_1}) \mid v_2;\theta_{iV_2})\Big] dv_1 dv_2.\end{aligned}$$

An approximation for the two-dimensional integral is given as

$$\begin{aligned}\pi_{1:d}(\mathbf{y},\theta) &\approx \sum_{k_1=1}^{n_q}\sum_{k_2=1}^{n_q} w_{k_1} w_{k_2} \prod_{i=1}^{d} \Big[C_{i|V_2;V_1}(F_{i|V_1}(y_i \mid v_1;\theta_{iV_1}) \mid v_2;\theta_{iV_2}) \\ &\qquad -C_{i|V_2;V_1}(F_{i|V_1}(y_i-1 \mid v_1;\theta_{iV_1}) \mid v_2;\theta_{iV_2})\Big] dv_1 dv_2.\end{aligned}$$

Algorithms (3) and (4) provide details on likelihood computations for the one- and two-factor copula models.

The maximum likelihood estimator $\hat{\theta}$ is obtained using numerical optimization via the routines in the R software. To ensure that the estimated parameters are within the correct parameter range for the bivariate linking

Algorithm 3 Likelihood for one factor copula

Likelihood for one factor copula with ordinal response and no covariates, with cutpoints based on univariate margins (Nikoloulopoulus and Joe(2014)). There are d ordinal variables, category labels $0, \ldots, k-1$ for each variable. V is the latent variable assumed to be U(0,1)

1. Input Q= number of quadrature points for Gauss-Legendre, the nodes x_q and the weights w_q, d = number of variables, K= number of categories for each variable n= sample size, d-vectors $y_1, ..., y_n$ with elements in $0, ..., K-1^d$, copula C_{V_i} (linking ith observed variable Y_i to the latent variable V) for $i = 1, ...d$ and their parameters values.
2. Set the cutpoints on the uniform scale $a_{11}, ..., a_{1,K-1}, ...a_{d1}, ..., a_{d,K-1}$ based on the univariate margins pmfs. Also set the boundary cutpoints $a_{i0} = 0$ and $a_{iK} = 1$ for each variable.
3. For $i = 1, ...d$ compute/store $C_{i|V}(a_{ik}|x_q)$ for $k = 0, ..., K$ and $q = 0, ..., Q$
4. For $i = 1, ...d$ compute the probability $f_{i|V}(k-1|x_q) = C_{i|V}(a_{ik}|x_q) - C_{i|V}(a_{i,k-1}|x_q)$ for $k = 0, ..., K$ and $q = 0, ..., Q$. After this step assume all these conditional pmfs are stored in $Q \times K \times d$ array
5. loglik $\leftarrow 0$
6. for $j = 1, ..., n$: (data loop) do
7. for y_j, let $p_{jq} \leftarrow \prod_{i=1}^{d} f_{i|V}(y_{ji}|x_q)$ by extracting from the 3-dimensional array
8. update the log-likelihood with loglik $\leftarrow$ loglik+log$(\sum_{q=1}^{Q} p_{jq} w_q)$
9. end for
10. Return loglik

copulas, we use reparameterizations suggested in Acar et al. (2011). For example, for the Gumbel copula parameter $\theta \in [1, \infty)$, we use $\eta = \log(\theta - 1)$ and perform unconstraint optimization for $\eta \in (-\infty, \infty)$.

As in the traditional factor model, in factor copula analysis, the factor

Algorithm 4 Likelihood for two factor copula

Likelihood for 2-factor copula with ordinal response and no covariates, with cutpoints based on univariate margins (Nikoloulopoulus and Joe(2014)). There are d ordinal variables, category labels $0, \ldots, k-1$ for each variable. V_1, V_2 are the latent variables assumed to be independent U(0,1)

1. Input Q= number of quadrature points for Gauss-Legendre, the nodes x_q and the weights w_q, d = number of variables, K= number of categories for each variable n= sample size, d-vectors $y_1, ..., y_n$ with elements in $0, ..., K-1^d$, copula $C_{V_{1_i}}$ (linking ith observed variable Y_i to the latent variable V_1) and $C_{V_2 i;V_1}$ (linking ith observed variable Y_i to the latent variable V_2 given V_1) for $i = 1, ...d$ and their parameters Values.
2. Set the cutpoints on the uniform scale $a_{11}, ..., a_{1,K-1}, ...a_{d1}, ..., a_{d,K-1}$ based on the univariate margins pmfs. Also set the boundary cutpoints $a_{i0} = 0$ and $a_{iK} = 1$ for each variable.
3. For $i = 1, ...d$ compute/store $s_{q_1,k,i} = C_{i|V}(a_{ik}|x_q)$ for $k = 0, ..., K$ and $q = 0, ..., Q$
4. For $i = 1, ...d$ compute/store $t_{q_1,q_2,k,i} = C_{i|V_2;V_1}(C_{i|V_1}(a_{ik}|x_{q_1}|x_{q_2})) = C_{i|V_2;V_1}(s_{q_1,k,i}|x_{q_2}$ for $k = 0, ..., K$ and $q_1, q_2 = 0, ..., Q$,
5. For $i = 1, ...d$ compute the probability $f_{i|V_1,V_2}(k-1|x_{q_1}, x_{q_2})) = t_{q_1,q_2,k,i} - t_{q_1,q_2,k-1,i}$ for for $k = 0, ..., K$ and $q_1, q_2 = 0, ..., Q$. After this step assume all these conditional pmfs are stored in $Q \times Q \times K \times d$ array
6. loglik $\leftarrow 0$
7. for $j = 1, ..., n$: (data loop) do
8. for y_j, let $p_{jq_1q_2} \leftarrow \prod_{i=1}^{d} f_{i|V_1,V_2}(y_{ji}|x_{q_1}, x_{q_2}))$ by extracting from the 4-dimensional array
9. update the log-likelihood with loglik $\leftarrow$ loglik+log$(\sum_{q=1}^{Q} p_{jq_1q_2} w_{q_1} w_{q_2})$
10. end for
11. Return loglik

scores are obtained using model-based predictions. Given the copula parameter estimates, we use the conditional expectation formula to obtain the expected factor scores. For a one-factor copula model, the factor scores given data are obtained using

$$\mathrm{E}(V|\mathbf{Y}) = \int_0^1 v \frac{\prod_{i=1}^d (C_{i|v}(a_{i,y_{i+1}}|v) - C_{i|v}(a_{i,y_i}|v))}{\int_0^1 \prod_{i=1}^d (C_{i|v}(a_{i,y_{i+1}}|v) - C_{i|v}(a_{i,y_i}|v))dv} dv, \tag{4.3}$$

where a's are the cut-points and the ratio inside the integral gives the conditional density of V given $\mathbf{Y}$. The integrals are calculated numerically via the Gauss-Legendre quadrature.

Similarly, for a two-factor copula model, the conditional expectations are defined as

$$\mathrm{E}(V_1|\mathbf{Y}) = \int_0^1 v_1 \frac{\int_0^1 \prod_{i=1}^d f_{iV_2|V_1}(v_2, y_i \mid v_1) dv_2}{\int_0^1 \int_0^1 \prod_{i=1}^d \Big[f_{iV_2|V_1}(v_2, y_i \mid v_1) \Big] dv_1 dv_2} dv_1,$$

$$\mathrm{E}(V_2|\mathbf{Y}) = \int_0^1 v_2 \frac{\int_0^1 \prod_{i=1}^d f_{iV_2|V_1}(v_2, y_i \mid v_1) dv_1}{\int_0^1 \int_0^1 \prod_{i=1}^d \Big[f_{iV_2|V_1}(v_2, y_i \mid v_1) \Big] dv_1 dv_2} dv_2,$$

where

$$f_{iV_2|V_1}(v_2, y \mid v_1) = C_{i|V_2;V_1}(C_{i|V_1}(a_{i,y_{i+1}}|v_1) \mid v_2) - C_{i|V_2;V_1}(C_{i|V_1}(a_{i,y_i} \mid v_1) \mid v_2).$$

4.3 Numerical Assessments

A Monte Carlo simulation study is designed to assess the robustness of factor scores estimation and regression. Our primary goal is to test associations between the latent variable and a covariate. Hence, we also assess the type I error rate and empirical statistical power of the polychoric correlation-based factor analysis and the factor copula based analysis.

Below we first outline the simulation setting and present the results on these aspects.

4.3.1 Simulation Setting

To generate data suitable for factor scores regression, we employ the algorithm for the two-factor copula model presented in Chapter 2.

We investigate the problem where the latent variables are associated with a covariate. The covariate data on X are generated for a single SNP with minor allele frequency (MAF) 10% from the multinomial distribution with probabilites $(0.81, 0.18, 0.01)$. The three cases considered for $n = 200, 500$ and 1000 resulted in the following empirical counts $(167, 30, 3)$, $(410, 78, 12)$ and $(816, 168, 16)$ respectively. Our aim is to test

$$\mathrm{H}_0 : \beta_1^{(\mathrm{p})} = 0 \quad \text{versus} \quad \mathrm{H}_\mathrm{A} : \beta_1^{(\mathrm{p})} \neq 0$$

for $p = 1, 2$. Note that the factors are orthogonal, hence the test can be performed separately.

When generating the latent factors Z_1 and Z_2 from the factor scores regression model, the values of β_1 are chosen so that the coefficient of determination is $Q = 0\%, 1\%$ and 5%. While $Q = 0\%$ corresponds to the null hypothesis of no association, $Q = 1\%$ and 5% allow us to evaluate the power for small and moderate covariate effects on the latent factors. The corresponding slope coefficients are $\beta = 0, 0.2$ and 0.5, respectively. Under each setting, we consider $\beta_1^{(1)} = \beta_1^{(2)}$, i.e., the same covariate effects on each factor.

For ordinal manifest variables, we consider six items, with the same loading matrix in (3.2.1). Hence, the first three items were explained by Z_1 and the next three were linked to Z_2. Since, autism severity is expected to show right-skewness, we focus on the marginal distribution $p_3 = (0.60, 0.20, 0.15, 0.05)$ in Table 3.1. As in the simulation setting in Chapter 3, we generate samples of size $n = 200, 500$ and 1000 under the Gaussian, Frank, Gumbel and Clayton linking copulas. Under each setting, the experiment is repeated 1000 times.

4.3.2 Results on Factor Analysis

For each simulated sample, we estimate the polychoric correlation matrix given the multivariate ordinal data and perform the traditional factor analysis using the R polycor package (Fox, 2007). Under the traditional factor analysis, in

order to improve the interpretability of the extracted factors, factor rotations are performed. The two common types are the orthogonal (varimax) rotation which has the restriction that the two factors are independent and oblique (oblimin) rotation where the factors can be correlated with each other. To determine whether the factor rotations affect the tests results, we estimate the factor scores under the varimax, oblimin and no rotation.

We first investigate the underlying factor structure in the dataset by checking the number of factors. To determine the number of factors to include in a factor model, the most commonly used approaches are scree plot, eigenvalue greater than one rule, minimum average partial test, parallel analysis among others (Horn, 1965; Kaiser, 1960; Revelle, 2017). In our implementation, we use the psych package (Revelle and Revelle, 2017) which is based on the parallel analysis of the polychoric correlation matrices. This technique is simulation-based and compares the eigenvalues computed from a random data matrix with the same number of variables and observations as the real dataset from the analysis. The estimated eigenvalues and the original eigenvalues are plotted against the factors and the number of factors retained are based on the number of factors before the intersection of the two curves (Horn, 1965). We also obtain the estimated loadings and estimated factor scores.

Table (4.1) presents the results on the number of factors selected over the 1000 runs under each setting.

Table 4.1: Number of runs (out of 1000) the k^{th} factor model is selected under each copula family, for $k = 2, \ldots, 5$.

		$\beta_1 = 0$				$\beta_1 = 0.2$				$\beta_1 = 0.5$			
Copula	n	k=2	k=3	k=4	k=5	k=2	k=3	k=4	k=5	k=2	k=3	k=4	k=5
Gaussian	200	949	46	5	-	950	47	3	-	956	44	-	-
	500	988	11	1	-	980	19	1	-	981	19	-	-
	1000	994	6	-	-	994	5	1	-	993	7	-	-
Frank	200	945	51	4	-	931	67	2	-	934	65	1	-
	500	978	22	-	-	977	22	1	-	978	22	-	-
	1000	995	4	1	-	993	6	1	-	991	9	-	-
Gumbel	200	984	16	-	-	983	17	-	-	987	13	-	-
	500	996	3	1	-	998	2	-	-	998	2	-	-
	1000	1000	-	-	-	991	1	-	-	1000	-	-	-
Clayton	200	791	179	29	1	793	183	23	1	778	204	18	
	500	880	109	11	-	882	105	13	-	870	120	10	-
	1000	938	56	6	-	937	59	4	-	942	55	3	-

As can be seen, the success probability of correct identification of the factor model is fairly high in the traditional factor analysis. From our results, the suggested factors are ranging from 2 to 5. The two-factor model is correctly identified roughly between 77.8% and 100% of the time. The second most selected model is the three-factor model with less than 20% of the time.

Despite that the number of factors selected can be different than two, we fit a two-factor model in each run, mainly for interpretability of loading matrix and factor scores estimates. The estimated loadings averaged over 1000 replications along with their standard errors under each copula family are displayed in Table 4.2 for the case where varimax rotation is used.

Table 4.2: Average estimated loadings under varimax rotation under each copula family based on 1000 Monte-Carlo samples

	Gaussian		Frank		Gumbel		Clayton	
Items	$\mathrm{E}(\hat{L})$	$S.E.(\hat{L})$	$\mathrm{E}(\hat{L})$	$S.E.(\hat{L})$	$\mathrm{E}(\hat{L})$	$S.E.(\hat{L})$	$\mathrm{E}(\hat{L})$	$S.E.(\hat{L})$
ℓ_{11}	0.629	0.084	0.625	0.081	0.612	0.076	0.632	0.096
ℓ_{21}	0.389	0.086	0.396	0.082	0.374	0.075	0.454	0.101
ℓ_{31}	0.168	0.022	0.185	0.022	0.162	0.019	0.199	0.030
ℓ_{41}	-0.007	0.031	-0.007	0.032	-0.007	0.039	-0.005	0.032
ℓ_{51}	-0.006	0.026	-0.006	0.026	-0.007	0.029	-0.006	0.029
ℓ_{61}	0.000	0.018	0.000	0.019	-0.001	0.019	0.001	0.022
ℓ_{12}	-0.007	0.033	-0.006	0.034	-0.008	0.040	-0.005	0.034
ℓ_{22}	-0.005	0.025	-0.005	0.026	-0.006	0.029	-0.006	0.027
ℓ_{32}	0.000	0.018	0.000	0.018	-0.001	0.018	0.001	0.021
ℓ_{42}	0.627	0.086	0.625	0.081	0.612	0.077	0.632	0.100
ℓ_{52}	0.391	0.087	0.396	0.082	0.374	0.075	0.453	0.104
ℓ_{62}	0.168	0.023	0.185	0.023	0.162	0.019	0.198	0.033

From the estimated loadings, each item loads to a distinct factor. The first

factor Z_1 explains about 17% to 63% of the variation in Y_1, Y_2 and Y_3 whiles the second factor explains about 16% to 74% of the variation in Y_4, Y_5 and Y_6. The signs of the loadings show the direction of the correlation and do not affect the interpretation of the factor loadings when the magnitude is small.

4.3.3 Results on Factor Copula Analysis

Under each setting, we fit a factor copula model using the correct copula family for the bivariate linking copulas and obtain the estimates of the dependence parameters by maximizing the likelihood in Algorithm (4). The estimation performance is assessed using the mean and standard errors, as well as the bias of the parameter estimates. These are obtained using

$$\widehat{\mathrm{E}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r, \qquad \text{and} \qquad \widehat{Bias}(\hat{\theta}) = \widehat{\mathrm{E}}(\hat{\theta}) - \theta,$$

$$S.E.(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r\right)^2}$$

where R is the number of non-divergent runs. The estimation results are summarized in Tables (4.3), (4.4) ,(4.5) and (4.6).

From the tables, we see that the parameter estimation is overall satisfactory, despite some divergence issues under the Clayton copula. In comparison

Table 4.3: Relative bias and standard error of the copula parameter estimates under the two-factor Gaussian copula model based on $R = 600$ Monte-Carlo samples.

Parameter	θ	$\mathrm{E}(\hat{\theta})$	$S.E.(\hat{\theta})$	Relative Bias
θ_{1,v_1}	0.900	0.896	0.030	-0.004
θ_{2,v_1}	0.800	0.802	0.032	0.003
θ_{3,v_1}	0.500	0.501	0.033	0.002
θ_{4,v_1}	0.010	0.016	0.038	0.600
θ_{5,v_1}	0.010	0.013	0.039	0.300
θ_{6,v_1}	0.010	0.016	0.043	0.600
$\theta_{1,v_2;v_1}$	0.023	0.028	0.039	0.217
$\theta_{2,v_2;v_1}$	0.017	0.017	0.046	0.000
$\theta_{3,v_2;v_1}$	0.012	0.012	0.048	0.000
$\theta_{4,v_2;v_1}$	0.900	0.896	0.031	-0.004
$\theta_{5,v_2;v_1}$	0.800	0.804	0.031	0.005
$\theta_{6,v_2;v_1}$	0.500	0.503	0.039	0.006

Table 4.4: Relative Bias and standard error of the copula parameter estimates under the two-factor Frank copula model based on $R = 1000$ Monte-Carlo samples.

Parameter	θ	$\mathrm{E}(\hat{\theta})$	$S.E.(\hat{\theta})$	Relative Bias
θ_{1,v_1}	12.025	11.921	1.108	-0.009
θ_{2,v_1}	7.677	7.726	0.751	0.006
θ_{3,v_1}	3.306	3.314	0.310	0.002
θ_{4,v_1}	0.057	0.096	0.334	0.684
θ_{5,v_1}	0.057	0.075	0.320	0.316
θ_{6,v_1}	0.057	0.082	0.285	0.439
$\theta_{1,v_2;v_1}$	0.131	0.165	0.790	0.260
$\theta_{2,v_2;v_1}$	0.096	0.106	0.533	0.104
$\theta_{3,v_2;v_1}$	0.066	0.077	0.334	0.167
$\theta_{4,v_2;v_1}$	12.029	11.995	1.213	-0.003
$\theta_{5,v_2;v_1}$	7.678	7.794	0.741	0.015
$\theta_{6,v_2;v_1}$	3.306	3.322	0.325	0.005

with the magnitude of the true values and considering the parameter range under each family, the bias is fairly small for most bivariate linking copu-

Table 4.5: Relative Bias and standard error of the copula parameter estimates under the two-factor Gumbel copula model based on $R = 1000$ Monte-Carlo samples.

Parameter	θ	$\mathrm{E}(\hat{\theta})$	$S.E.(\hat{\theta})$	Relative Bias
θ_{1,v_1}	3.483	3.624	0.826	0.040
θ_{2,v_1}	2.441	2.473	0.222	0.013
θ_{3,v_1}	1.500	1.502	0.057	0.001
θ_{4,v_1}	1.006	1.010	0.010	0.004
θ_{5,v_1}	1.006	1.009	0.009	0.003
θ_{6,v_1}	1.006	1.010	0.012	0.004
$\theta_{1,v_2;v_1}$	1.015	1.025	0.031	0.010
$\theta_{2,v_2;v_1}$	1.011	1.018	0.020	0.007
$\theta_{3,v_2;v_1}$	1.007	1.011	0.013	0.004
$\theta_{4,v_2;v_1}$	3.484	3.691	1.298	0.059
$\theta_{5,v_2;v_1}$	2.441	2.480	0.222	0.016
$\theta_{6,v_2;v_1}$	1.500	1.503	0.058	0.002

Table 4.6: Relative Bias and standard error of the copula parameter estimates under the two-factor Clayton copula model based on $R = 993$ Monte-Carlo samples.

Parameter	θ	$\mathrm{E}(\hat{\theta})$	$S.E.(\hat{\theta})$	Relative Bias
θ_{1,v_1}	4.965	5.631	4.103	0.134
θ_{2,v_1}	2.882	3.057	0.764	0.061
θ_{3,v_1}	1.000	1.012	0.419	0.012
θ_{4,v_1}	0.013	0.016	0.017	0.231
θ_{5,v_1}	0.013	0.014	0.013	0.077
θ_{6,v_1}	0.013	0.015	0.017	0.154
$\theta_{1,v_2;v_1}$	0.030	0.041	0.053	0.367
$\theta_{2,v_2;v_1}$	0.021	0.026	0.034	0.238
$\theta_{3,v_2;v_1}$	0.015	0.016	0.013	0.067
$\theta_{4,v_2;v_1}$	4.967	5.630	3.959	0.133
$\theta_{5,v_2;v_1}$	2.883	3.152	2.701	0.093
$\theta_{6,v_2;v_1}$	1.000	1.013	0.157	0.013

las. Furthermore, the standard errors under most linking copulas are also very small indicating a relative precision and consistency of the parameter

estimates. Overall, the MLE provides quite accurate estimates of the factor copula parameters for each bivariate linking copula with a maximum bias of 0.785 under the Clayton copula.

4.3.4 Results on Factor Scores Regression

For each sample, the fitted traditional two-factor models with no rotation, and varimax and oblimin rotations and the fitted factor copula models are used to estimate factor scores. Under the traditional factor model, the estimated factor scores are in the normal scale. For the factor copula model, we obtain the expected scores in uniform scale and converted them to the normal scale using the inverse cdf of the standard normal distribution.

Given the estimated factor scores, we fit the simple linear model in (2.2) and perform a t-test to investigate associations of each factor with the covariate X. We obtain the estimates of the regression effects and the p-values.

To assess the robustness of the factor scores regression, we compute the empirical type-I error under the null hypothesis ($\beta_1 = 0$) and the empirical power under the alternative hypothesis $\beta_1 = 0.2$ and $\beta_1 = 0.5$. The empirical rejection rate is calculated as the proportion of the simulated datasets for which p-value is less than $\alpha = 0.01$ or 0.05, i.e.,

$$\frac{\sum_{r=1}^{R} \mathbb{I}(p\text{-value} \leq \alpha)}{R},$$

where $\mathbb{I}$ denotes the indicator function.

The robustness of the factor score regression methods is assessed by comparing the estimates with the true scores (benchmark). The robustness of the type 1 error is assessed based on the liberal criterion (Bradley, 1978) and the binomial test. When the null hypothesis is true, then the true probability ρ of a type 1 error is equivalent to the significance level α. Bradley (1978) proposed a quantitative definition of robustness where for a given α value, a range of ρ values are computed for which the test would be robust. The liberal criterion is based on $0.5\alpha \leq \rho \leq 1.5\alpha$. That is for all cases where the ρ values are within that interval, the test would be considered as robust.

As an alternative way of assessing robustness of the type I error estimates, we also consider the binomial distribution for the rejection rates under the null hypothesis and check whether the reported estimates differ significantly from the nominal level α. For this, we calculate $\alpha \pm 2\sqrt{\alpha(1-\alpha)/M}$, where $M = 1000$.

Table (4.7), (4.8), (4.9) and (4.10) summarizes the estimates of empirical rejection rates for the factor score regression and the factor copula scores regression with Gaussian, Gumbel, Clayton and Frank bivariate linking copulas.

For $\alpha = 0.01$ and $\alpha = 0.05$, the confidence interval under the binomial test is $(0.0037, 0.0163), (0.0362, 0.0638)$ respectively whiles based on the liberal criterion the confidence intervals for $\alpha = 0.01$ and $\alpha = 0.05$ are $(0.005, 0.015)$ and $(0.025, 0.075)$, respectively. In the case of no effect ($\beta_1 = 0$), we see that

based on the liberal criterion and the binomial test, all the recorded values are within the interval. The estimates of the type 1 error rates are not significantly different from the truth. All the methods yield approximately the correct type 1 error. Hence, there is no substantial difference in the type 1 error rates under violations of the distributional assumptions.

The latency of the factor leads to a power loss under all the scenarios considered. It is difficult to assess the significance of the drop in the power. We therefore mainly evaluate the nominal values that are reported. For instance, if the power is 1 and it drops to 0.98, depending on the context, that may not be considered as a large drop. For small effect size ($\beta_1 = 0.2$), there is a considerable loss of power, especially under no rotation. However, the traditional factor analysis with the oblimin and varimax rotations performed reasonably well. Factor rotations results in a more accurate identification of the factors and yield more accurate test results. We get similar results under the factor copula model.

When the regression coefficient is moderate ($\beta_1 = 0.5$), we see that all three approaches are pretty close, whereas no rotation suffers more from the power loss. The discrepancy between the benchmark values and the varimax, oblimin and factor copula methods becomes less significant as almost all yield 100% rejection. As expected, the empirical power increases with the sample size. Overall, the factor copula offers a slight power gain for the scenarios considered.

Table 4.7: Empirical rejection rates under the Gaussian copula

n	Method		$\beta_1 = 0$		$\beta_1 = 0.2$		$\beta_1 = 0.5$	
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
200	**Benchmark**	F_1	0.008	0.039	0.102	0.251	0.697	0.874
		F_2	0.009	0.055	0.097	0.250	0.678	0.863
	No Rotation	F_1	0.006	0.040	0.108	0.213	0.490	0.602
		F_2	0.012	0.052	0.088	0.196	0.448	0.567
	Varimax	F_1	0.004	0.046	0.082	0.212	0.545	0.744
		F_2	0.004	0.054	0.080	0.212	0.539	0.725
	Oblimin	F_1	0.003	0.047	0.083	0.217	0.582	0.775
		F_2	0.004	0.057	0.086	0.217	0.572	0.755
	Factor copula	F_1	0.008	0.040	0.078	0.221	0.566	0.775
		F_2	0.005	0.060	0.084	0.208	0.513	0.727
500	**Benchmark**	F_1	0.009	0.049	0.365	0.625	0.997	1.000
		F_2	0.012	0.050	0.385	0.619	0.994	0.998
	No Rotation	F_1	0.009	0.042	0.288	0.443	0.580	0.634
		F_2	0.009	0.059	0.303	0.465	0.599	0.650
	Varimax	F_1	0.008	0.046	0.277	0.489	0.970	0.992
		F_2	0.012	0.050	0.278	0.509	0.969	0.989
	Oblimin	F_1	0.008	0.047	0.281	0.496	0.980	0.993
		F_2	0.012	0.054	0.294	0.525	0.978	0.996
	Factor copula	F_1	0.009	0.050	0.279	0.495	0.990	1.000
		F_2	0.011	0.052	0.285	0.514	0.970	0.995
1000	**Benchmark**	F_1	0.008	0.046	0.714	0.876	1.000	1.000
		F_2	0.010	0.056	0.680	0.867	1.000	1.000
	No Rotation	F_1	0.008	0.044	0.450	0.565	0.600	0.654
		F_2	0.008	0.070	0.434	0.539	0.583	0.640
	Varimax	F_1	0.009	0.059	0.540	0.748	1.000	1.000
		F_2	0.005	0.059	0.502	0.715	0.999	0.999
	Oblimin	F_1	0.009	0.057	0.561	0.760	1.000	1.000
		F_2	0.006	0.059	0.525	0.727	1.000	1.000
	Factor copula	F_1	0.010	0.060	0.55	0.763	1.000	1.000
		F_2	0.010	0.050	0.485	0.713	1.000	1.000

Table 4.8: Empirical rejection rates under the Frank copula

n	Method		$\beta_1 = 0$		$\beta_1 = 0.2$		$\beta_1 = 0.5$	
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
200	**Benchmark**	F_1	0.008	0.039	0.102	0.251	0.697	0.874
		F_2	0.009	0.055	0.097	0.250	0.678	0.863
	No Rotation	F_1	0.008	0.060	0.094	0.205	0.472	0.589
		F_2	0.007	0.042	0.081	0.187	0.413	0.550
	Varimax	F_1	0.009	0.048	0.081	0.205	0.515	0.713
		F_2	0.006	0.050	0.077	0.195	0.490	0.701
	Oblimin	F_1	0.009	0.046	0.083	0.211	0.548	0.735
		F_2	0.006	0.052	0.080	0.205	0.527	0.744
	Factor copula	F_1	0.008	0.050	0.075	0.203	0.510	0.734
		F_2	0.006	0.053	0.078	0.185	0.493	0.693
500	**Benchmark**	F_1	0.009	0.049	0.365	0.625	0.997	1.000
		F_2	0.012	0.050	0.385	0.619	0.994	0.998
	No Rotation	F_1	0.008	0.054	0.274	0.420	0.569	0.624
		F_2	0.009	0.052	0.274	0.449	0.598	0.657
	Varimax	F_1	0.009	0.050	0.260	0.462	0.960	0.991
		F_2	0.013	0.052	0.249	0.477	0.590	0.985
	Oblimin	F_1	0.008	0.048	0.268	0.475	0.971	0.993
		F_2	0.012	0.048	0.268	0.487	0.965	0.992
	Factor copula	F_1	0.007	0.044	0.213	0.457	0.961	0.990
		F_2	0.012	0.053	0.263	0.463	0.955	0.981
1000	**Benchmark**	F_1	0.008	0.046	0.714	0.876	1.000	1.000
		F_2	0.010	0.056	0.680	0.867	1.000	1.000
	No Rotation	F_1	0.009	0.048	0.449	0.561	0.620	0.674
		F_2	0.011	0.066	0.426	0.526	0.564	0.621
	Varimax	F_1	0.010	0.054	0.497	0.718	0.997	0.999
		F_2	0.010	0.060	0.467	0.708	0.998	1.000
	Oblimin	F_1	0.010	0.054	0.508	0.731	0.999	1.000
		F_2	0.010	0.058	0.481	0.719	0.999	1.000
	Factor copula	F_1	0.007	0.060	0.499	0.747	1.000	1.000
		F_2	0.013	0.059	0.468	0.708	1.000	1.000

Table 4.9: Empirical rejection rates under the Gumbel copula

n	Method		$\beta_1 = 0$		$\beta_1 = 0.2$		$\beta_1 = 0.5$	
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
200	**Benchmark**	F_1	0.008	0.039	0.102	0.251	0.697	0.874
		F_2	0.009	0.055	0.097	0.250	0.678	0.863
	No Rotation	F_1	0.010	0.045	0.109	0.204	0.475	0.591
		F_2	0.011	0.050	0.096	0.198	0.479	0.575
	Varimax	F_1	0.012	0.046	0.081	0.220	0.567	0.770
		F_2	0.005	0.060	0.099	0.210	0.564	0.731
	Oblimin	F_1	0.012	0.049	0.089	0.224	0.605	0.798
		F_2	0.006	0.059	0.102	0.219	0.603	0.769
	Factor copula	F_1	0.008	0.043	0.094	0.213	0.556	0.775
		F_2	0.006	0.058	0.086	0.207	0.513	0.727
500	**Benchmark**	F_1	0.009	0.049	0.365	0.625	0.997	1.000
		F_2	0.012	0.050	0.385	0.619	0.994	0.998
	No Rotation	F_1	0.009	0.051	0.301	0.446	0.566	0.627
		F_2	0.011	0.054	0.316	0.466	0.582	0.632
	Varimax	F_1	0.008	0.053	0.288	0.509	0.977	0.995
		F_2	0.014	0.048	0.300	0.516	0.974	0.990
	Oblimin	F_1	0.008	0.055	0.302	0.522	0.985	0.994
		F_2	0.014	0.047	0.311	0.530	0.981	0.995
	Factor copula	F_1	0.008	0.049	0.287	0.501	0.977	0.994
		F_2	0.012	0.053	0.299	0.525	0.974	0.991
1000	**Benchmark**	F_1	0.008	0.046	0.714	0.876	1.000	1.000
		F_2	0.010	0.056	0.680	0.867	1.000	1.000
	No Rotation	F_1	0.005	0.039	0.464	0.565	0.570	0.631
		F_2	0.009	0.067	0.451	0.559	0.590	0.645
	Varimax	F_1	0.007	0.050	0.558	0.745	0.999	0.999
		F_2	0.008	0.055	0.520	0.719	0.999	1.000
	Oblimin	F_1	0.008	0.052	0.581	0.764	1.000	1.000
		F_2	0.006	0.055	0.550	0.737	1.000	1.000
	Factor copula	F_1	0.008	0.053	0.565	0.767	1.000	1.000
		F_2	0.007	0.052	0.521	0.737	1.000	1.000

Table 4.10: Empirical rejection rates under the Clayton copula

n	Method		$\beta_1 = 0$		$\beta_1 = 0.2$		$\beta_1 = 0.5$	
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
200	**Benchmark**	F_1	0.008	0.039	0.102	0.251	0.697	0.874
		F_2	0.009	0.055	0.097	0.250	0.678	0.863
	No Rotation	F_1	0.010	0.036	0.064	0.186	0.371	0.532
		F_2	0.008	0.043	0.065	0.160	0.382	0.526
	Varimax	F_1	0.005	0.051	0.065	0.166	0.390	0.622
		F_2	0.007	0.048	0.055	0.165	0.388	0.605
	Oblimin	F_1	0.004	0.049	0.066	0.172	0.427	0.649
		F_2	0.007	0.051	0.055	0.171	0.410	0.631
	Factor copula	F_1	0.010	0.044	0.057	0.173	0.407	0.649
		F_2	0.011	0.050	0.057	0.165	0.392	0.625
500	**Benchmark**	F_1	0.009	0.049	0.365	0.625	0.997	1.000
		F_2	0.012	0.050	0.385	0.619	0.994	0.998
	No Rotation	F_1	0.008	0.054	0.220	0.383	0.607	0.674
		F_2	0.008	0.051	0.226	0.387	0.617	0.678
	Varimax	F_1	0.010	0.050	0.193	0.398	0.909	0.969
		F_2	0.012	0.056	0.134	0.301	0.668	0.821
	Oblimin	F_1	0.007	0.052	0.198	0.402	0.904	0.965
		F_2	0.016	0.051	0.211	0.408	0.914	0.971
	Factor copula	F_1	0.010	0.052	0.211	0.433	0.921	0.981
		F_2	0.012	0.053	0.223	0.437	0.941	0.986
1000	**Benchmark**	F_1	0.008	0.046	0.714	0.876	1.000	1.000
		F_2	0.010	0.056	0.680	0.867	1.000	1.000
	No Rotation	F_1	0.009	0.057	0.383	0.533	0.665	0.724
		F_2	0.016	0.062	0.385	0.525	0.614	0.682
	Varimax	F_1	0.015	0.054	0.399	0.640	0.997	1.000
		F_2	0.008	0.063	0.402	0.613	0.995	0.998
	Oblimin	F_1	0.015	0.056	0.422	0.647	0.998	1.000
		F_2	0.009	0.061	0.418	0.618	0.998	0.999
	Factor copula	F_1	0.012	0.054	0.450	0.675	1.000	1.000
		F_2	0.016	0.058	0.414	0.667	1.000	1.000

Chapter 5

Conclusion

In the analysis of multivariate ordinal variables, factor analysis with polychoric correlations are commonly used to identify the underlying latent factors and to quantify the scores. In this thesis, we studied the dependence measures for multivariate ordinal data and investigated the robustness of the factor analysis to violations of distributional assumptions. We also investigated associations between the estimated factor scores and a given covariate using factor scores regression.

We addressed these aspects in a flexible framework called factor copulas which can accommodate a wide range of dependence patterns to generate multivariate ordinal data. For illustrations we considered six items with two latent factors underlying their dependence. Given the multivariate ordinal data, we estimated the polychoric correlation matrix and performed the traditional fac-

tor analysis. We also fitted a factor copula model and obtained estimates of dependence parameters. Under both approaches, we estimated the factor scores and performed factor scores regression.

Our results from the simulations indicated that polychoric correlations for likert scale variables under symmetric dependence and/or probability mass functions are very close to the Pearson correlation for continuous latent variables when quantifying dependence in multivariate ordinal data. For asymmetric dependence, we observed that the polychoric correlation loses its advantage, and the discrepancy between the Pearson correlations and the polychoric correlation increases as the degree of skewness in the marginal distributions.

The use of the factor score estimates results in power loss in association testing under both the traditional factor analysis and the factor copula analysis in comparison to the case with the known factors. The relative power loss is higher when the effect size is small.

In this study, we assumed that the copula family is known for each bivariate linking copula. In practice, one needs to decide on which copula family best suits the data. As suggested in Nikoloulopoulos and Joe (2015), for a given dataset, different parametric copula families can be compared via the log-likelihood or Akaike information criterion (AIC). The model that gives the largest likelihood or smallest AIC will be chosen.

Our investigations were limited to two-factor copula model. As the number of factors increases, fitting a factor copula model becomes notoriously difficult.

Hence, despite their generality, factor copula models offer a very limited advantage in practice. Their use in data generation and assessment of robustness however, can be extended to factor models with higher number of factors.

The scenarios considered in this thesis were under the ideal case where all the latent variables are normally distributed. If the latent factors have skewed distributions (e.g., log-normal) the traditional factor analysis approach is expected to fail drastically. Future research is needed to examine such settings. Further considerations include situations where the items have different marginal distribution (e.g., some skewed and some symmetric) and different copula families in their relation to the factors.

Bibliography

Acar, E. F., Craiu, R. V., and Yao, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics*, 67(2):445–453.

Alan, A. (2010). *Analysis of Ordinal Categorical Data*. Probability and Statistics. Wiley.

Association, A. P. et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Bai, A., Hira, S., and Deshpande, P. (2015). An application of factor analysis in the evaluation of country economic rank. *Procedia Computer Science*, 54:311–317.

Balakrishnan, N. and Lai, C.-D. (2009). *Continuous bivariate distributions*. Springer Science & Business Media.

Bollen, K. A. and H.Barb, K. (1981). Pearson's r and coarsely categorized measures. *American Sociologilcal Review*, 46(2):232–239.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2):144–152.

Bralten, J., van Hulzen, K., Martens, M., Galesloot, T., Arias Vasques, A., Kimeney, L., Buitelaar, J., Muntjewerf, J., Franke, B., and Poelmans, G. (2017). Autism spectrum disorders and autistic traits share genetics and biology. *Molecular Psychiatry*, pages 1–8.

Brown, A., Ding, Z., Viñuela, A., Glass, D., Parts, L., Spector, T., Winn, J., and Durbin, R. (2015). Pathway based factor analysis of gene expression data produces highly heritable phenotypes that associate with age. *G3: Genes— Genomes— Genetics*, pages g3–114.

Camp, B. H. (1933a). Karl pearson and mathematical statistics. *Journal of the American Statistical Association*, 28(184):395–401.

Camp, B. H. (1933b). Karl pearson and mathematical statistics. *Journal of the American Statistical Association*, 28(184):395–401.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Cavallini, M., Di Bella, D., Sliprandi, F., Malciodi, F., and Bellodi, L. (2002). Exploratory factor analysis of obsessive-compulsive patients and associa-

tion with 5-httlpr polymorphism. *American Journal of Medical Genetics*, 114:347–353.

Chaste, P., Leboyer, M., et al. (2012). Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin Neurosci*, 14(3):281–92.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93:40–57.

Ekstrom, J. (2008). *On the relation between the phi-coefficient and the tetrachoric correlation coefficient*. In contributions to the theory of measures of association for ordinal variables, Acta Universitatis Upsaliensis.

Ekström, J. (2011). A generalized definition of the polychoric correlation coefficient.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272.

Fox, J. (2007). Polycor: polychoric and polyserial correlations. *R package version 0.7-5, URL http://CRAN. R-project. org/package= polycor*.

Francis K. C. Hui, David I. Warton, J. T. O. V. H. and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26(1):35–43.

Francisco P.Holgado-Tello, S. C.-M., Barbero-Garcia, I., and Villa-Abad, E. (2008). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological methods*, 6(4):430.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago Press.

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(3):497–508.

Hershberger, S. L. (2005). Factor score estimation. *Encyclopedia of statistics in behavioral science*.

Hewson, P. (2015). Cran task view: Multivariate statistics.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Ines Devlieger, Axel Mayer, Y. R. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement.*, 76(5):741–770.

Jin, S. and Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. *psychometrika*, 82(1):67–85.

Joakim, E. (2008). *A generalized definition of the polychoric correlation coefficient*. PhD thesis, Acta Universitatis Upsaliensis.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability,V.133. CRC Press.

Johnson, R. A., Wichern, D. W., et al. (2014). *Applied multivariate statistical analysis*, volume 4. Prentice-Hall New Jersey.

Joreskog, K. G. and Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Beavioral Research*, 36(3):347–387.

Jussila, K. K., Lyall, K., Kuusikko-Gauffin, S., Mattila, M.-L., Pollock-Wurman, R., Hurtig, T., Joskitt, L., Bloigu, R., Ebling, H., Moilanen, I., et al. (2015). Familiality of quantitative autism traits. *Scandinavian Journal of Child and Adolescent Psychiatry and Psychology*, 3(2):126–135.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Kaplunovsky, A. S. (2005). Factor analysis in environmental studies. *HAIT J. Sci. Eng. B*, 2(1-2):54–94.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Klüppelberg, C. and Kuhn, G. (2009). Copula structure analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):737–753.

Krupskii, P. (2014). *Structured factor copulas and tail inference*. PhD thesis, University of British Columbia.

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

Landa, R. J. (2008). Diagnosis of autism spectrum disorders in the first 3 years of life. *Nature Clinical Practice Neurology*, 4(3):138–147.

Ledesma, R. D. and Valero-Mora, P. (2007). Determining the number of factors to retain in efa: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research & evaluation*, 12(2):1–11.

Li, M., Boehnke, M., Abecasis, G. R., and Song, P. X.-K. (2006). Quantitative trait linkage analysis using gaussian copulas. *Genetics*, 173(4):2317–2327.

Liu, X.-Q., Georgiades, S., Duku, E., Thompson, A., Devlin, B., Cook, E. H., Wijsman, E. M., Paterson, A. D., and Szatmari, P. (2011). Identification of genetic loci underlying the phenotypic constructs of autism spectrum disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(7):687–696.

Luo, H. (2011). *Analysis of ordinal variables using rank-based polychoric correlation*. PhD thesis, Uppsala Universitet.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56:337–357.

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

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Publishing Company, Incorporated, second edition edition.

Neslehova, J. (2004). *Dependence of Non-Continuous Random Variables*. PhD thesis, Carl von Ossietzky Universitat Oldenburg.

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

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, 14(4):485–500.

Page, J., Constantino, J. N., Zambrana, K., Martin, E., Tunc, I., Zhang, Y., Abbacchi, A., and Messinger, D. (2016). Quantitative autistic trait measurements index background genetic risk for asd in hispanic families. *Molecular Autism*, 7(1):39.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195:1–405.

Preacher, K. J. and MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior genetics*, 32(2):153–161.

Revelle, W. (2011). An overview of the psych package. *Department of Psychology Northwestern University. Accessed on March*, 3:2012.

Revelle, W. (2014). psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, Illinois*, 165.

Revelle, W. (2017). How to: Use the psych package for factor analysis and data reduction.

Revelle, W. and Revelle, M. W. (2017). Package 'psych'.

Rutter, M. (2000). Genetic studies of autism: from the 1970s into the millennium. *Journal of Abnormal Child Psychology*, 28(1):3–14.

Rutter, M. (2005). Incidence of autism spectrum disorders: changes over time and their meaning. *Acta paediatrica*, 94(1):2–15.

Salkind, N. J. (2007). *Encyclopedia of measurement and statistics*, volume 1. Sage.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

Szatmari, P. (1999). Heterogeneity and the genetics of autism. *Journal of Psychiatry and Neuroscience*, 24(2):159.

Team, R. C. (2017). R: A language and environment for statistical computing [internet]. vienna, austria; 2014.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327.