Njål Foldnes & Steffen Grønneberg (2019) Pernicious Polychorics: The Impact and Detection of Underlying Non-normality, *Structural Equation Modeling: A Multidisciplinary Journal,* DOI: 10.1080/10705511.2019.1673168

Pernicious polychorics: The impact and detection of underlying non-normality

Njål Foldnes and Steffen Grønneberg

Department of Economics

BI Norwegian Business School

Oslo, Norway 0484

Correspondence concerning this article should be sent to *njal.foldnes@bi.no*

Abstract

Ordinal data in social science statistics are often modeled as discretizations of a multivariate normal vector. In contrast to the continuous case, where SEM estimation is also consistent under non-normality, violation of underlying normality in ordinal SEM may lead to inconsistent estimation. In this article, we illustrate how underlying non-normality induces bias in polychoric estimates and their standard errors, and it is noteworthy that this bias is strongly affected by how we discretize. It is therefore important to consider tests of underlying multivariate normality. In this study we propose a parametric bootstrap test for this purpose. Its performance relative to the test of Maydeu-Olivares is evaluated in a Monte Carlo study. At realistic sample sizes, the bootstrap exhibited substantively better Type I error control and power than the Maydeu-Olivares test in ordinal data with ten dimensions or higher. R code for the bootstrap test is provided.

Keywords: ordinal data, structural equation modeling, polychoric correlation, parametric bootstrap

Pernicious polychorics: The impact and detection of underlying non-normality

Introduction

Ordinal data, such as responses to questionnaires, are common in the behavioral, educational and psychological sciences. Instead of neglecting the categorical nature of the data by applying methods developed for continuous data, methodology has been developed in an effort to model ordinal categorical data (e.g. Christoffersson, 1975; Muthén, 1984). In the context of covariance modeling like confirmatory factor analysis (CFA) and structural equation modeling (SEM), these methods have been found to outperform methods that assume continuous variables (Li, 2016b). The methodology is based on the assumption that the observed ordinal data represent unobserved continuous variables that have been discretized. That is, observed p-dimensional vectors X of ordinal observations are thought to be generated through the discretization of an unobserved, continuous p-dimensional random vector ξ with correlation matrix Σ. This model class includes ordinal SEM and the special case of ordinal CFA, as well as IRT models, see Takane & De Leeuw (1987) and Foldnes & Grønneberg (2019, Appendix A). In order to identify the polychoric correlation matrix, that is, the correlation matrix of ξ, it is typically assumed that ξ has a multivariate normal distribution (Pearson, 1901). Under this assumption Σ may be estimated with normal-theory based maximum likelihood (Olsson, 1979). Then we can fit a CFA/SEM model to the matrix of polychoric correlations using diagonally or unweighted least squares estimation. The methodology is implemented in current software packages like EQS (Bentler, 2006), Mplus (Muthén & Muthén, 2012), LISREL (Jöreskog & Sörbom, 2015) and lavaan (Rosseel, 2012), and is in frequent use in empirical research.

When observing a random sample from a continuous distribution, the underlying covariance matrix can always be estimated consistently using the empirical covariance matrix. Such a universally valid estimator does not exist in the ordinal case, and the important polychoric correlations of Olsson (1979) specifically assume that ξ is multivariate normal, and need not be a consistent estimator for Σ outside of this condition.

In practice, empirical research studies do not usually report on the tenability of the normality assumption. This may partly be due to the lack of statistical tests for underlying non-normality in popular software packages. Another factor is that several highly influential Monte Carlo studies such as Flora & Curran (2004) have suggested that moderate deviations from normality in ξ will not lead to substantive bias in the estimates of polychoric correlations (see Grønneberg & Foldnes, 2019, for more references and a fuller discussion on this point).

Most simulation studies on the robustness of polychoric correlations simulate ordinal data using the following two-step procedure. Firstly, a continuous random vector with a fixed covariance matrix Σ is generated using the Vale-Maurelli (Vale & Maurelli, 1983) simulation approach. Secondly, this continuous vector is discretized in a manner made precise below (see eq. (3)). Surprisingly, it turns out that this second step usually results in an ordinal random vector which is numerically equal to a discretized version of a multivariate normal vector with a correlation matrix slightly different from Σ. That is, the simulated data is such that it could have been generated by simulating from an exactly multivariate normal random vector and then discretized. As a result, very few simulation studies have in fact studied real violation of the underlying normality assumption in ordinal SEM. This was recently shown in Grønneberg & Foldnes (2019), based on the result that the Vale-Maurelli vector usually has a normal copula, as shown in Foldnes & Grønneberg (2015). The extent to which non-normality in ξ leads to bias in the polychoric coefficients and their standard errors, and consequently to invalid model inference is therefore partially an open question.

We will illustrate that non-normality in ξ may entail substantial estimation bias for the polychoric correlation. We expect such bias to propagate to estimates of parameters in the SEM/CFA model, since these models are fitted to the polychoric correlation matrix. Our illustrations also show that there are non-normal conditions where the bias of the polychoric correlations are minimal, but where standard inference procedures based on a

normal assumption are invalid, e.g. by producing incorrect standard errors of the estimates. We therefore expect that SEM inference as well as chi-square statistics for testing correct model specification become invalid as a result. Hence, it is important to develop and evaluate tests for underlying normality in ordinal datasets. Should such a test reject the normality hypothesis, we must interpret our estimated model with more caution. We suggest that researchers should therefore run tests for underlying normality and report the result of such tests in their studies.

In the lowest possible dimension, e.g. the bivariate case, several tests have been proposed and evaluated for testing discretized non-normality (Maydeu-Olivares et al., 2009; Jöreskog, 2005). In the present article our focus is on testing discretized normality for vectors of arbitrary dimensions. While there are tests for combining pairwise normality tests for all bivariate marginals (Raykov & Marcoulides, 2015), to the best of our knowledge, only one test has been proposed for directly testing normality in arbitrary dimensions (Maydeu-Olivares, 2006) but has still not been empirically evaluated, with the exception of a small study in Maydeu-Olivares (2006).

This article is organized as follows. We first illustrate that the polychoric estimator may be severely biased when the underlying normality assumption is violated. We then review the general discretized normality model and the test of Maydeu-Olivares (2006). A new parametric bootstrap test for underlying non-normality is then introduced. Next, we describe our simulation design, with an emphasis on proper simulation methodology for ordinal covariance models. In the non-normal simulation conditions, we assess the bias of the normal-theory polychoric correlations and their associated standard errors. The simulation results are then presented and discussed. The final section provides some concluding remarks.

Illustrations of the pernicious influence of underlying non-normality on normal theory polychoric correlations

In this section we provide exact calculations for the normality-based polychoric correlation in a specific bivariate non-normal case. Our aim is to demonstrate that underlying non-normality may result in severe bias in the population values of the polychoric correlations. This investigation expands similar studies, such as those by Monroe (2018) and Jin & Yang-Wallentin (2017). In contrast to previous studies, we follow Foldnes & Grønneberg (2019) and only consider underlying distributions with standard normal marginals. We find that the size and direction of this bias varies greatly as a function of the number and placements of the thresholds, and is difficult to predict. This strongly motivates testing for underlying normality.

Our illustration is restricted to the bivariate case. Since the polychoric correlation matrix is computed using bivariate distributions only, it suffices to study the bivariate case. A further advantage of restricting attention to the bivariate case is that we are able to visualize the non-normal distributions we work with. We note that the study reported in this section has a focus on a limited case, and should be followed up by further research. Its main limitation is that we mainly focus on a specific deviation from normality. We first provide a detailed introductory example which shows that polychoric correlations may have substantial bias, in both positive and negative directions. We then expand this by systematically varying the set up of the introductory illustrative example, and we study how this influences the bias of the polychoric correlations.

A simple illustration of polychoric bias

We here provide a case that incorporates both substantive positive and negative bias with respect to the true correlation among ξ_1 and ξ_2, the coordinates of ξ. The calculations are available as R (R Core Team, 2018) code in the supplementary online material. We use the concept of a *bivariate copula* to construct a non-normal bivariate distribution. For a

general overview on copulas, see Nelsen (2007) and Joe (1997). A bivariate copula $C(u, v)$ is the joint cumulative distribution function (CDF) of two random variables, each of which is uniformly distributed on $[0, 1]$. Sklar (1959) showed that for any bivariate vector (ξ_1, ξ_2) with continuous marginals, the joint CDF $H(a, b) = P(\xi_1 \leq a, \xi_2 \leq b)$ may be uniquely decomposed into its marginals and its copula C:

$$H(a, b) = C(F_1(a), F_2(b))$$

where F_1 and F_2 are the CDFs of ξ_1 and ξ_2, respectively.

In order to simulate from H, one may first simulate a random vector (U_1, U_2) from the copula using general techniques described e.g. in Joe (1997), and then transform (U_1, U_2) via

$$(\xi_1, \xi_2) = \left(F_1^{-1}(U_1), F_2^{-1}(U_2)\right), \tag{1}$$

which will have distribution H, i.e. (ξ_1, ξ_2) will have marginal distributions F_1 and F_2, and copula C.

Following arguments in Foldnes & Grønneberg (2019), we here assume that both ξ_1 and ξ_2 are standard normally distributed, i.e. that $F_1 = F_2 = \Phi$, where Φ is the CDF for standard normal distribution. Given any copula C, a bivariate distribution with standard normal marginals is defined by the joint CDF $H(a, b) = C(\Phi(a), \Phi(b))$. In the present illustration we let C belong to the class of Joe copulas:

$$C_\theta(u, v) = 1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta}.$$

The strength of dependency is parameterized by $\theta \in [1, \infty]$. In this illustration the Pearson correlation between ξ_1 and ξ_2 is fixed at $\rho = 0.7$. A numerical search revealed that setting $\theta = 3.011$ for the Joe copula results in a Pearson correlation of $\rho = 0.7$ when combined with standard normal marginals. That is, the bivariate vector whose CDF is

$$H(a, b) = C_{\theta=3.011}(\Phi(a), \Phi(b)) \tag{2}$$

has a Pearson correlation coefficient of 0.7, and standard normal marginal distributions. In terms of the stochastic representation in eq.(1) with $F_1 = F_2 = \Phi$, we have that

$$0.7 = \mathrm{Cov}\left(\xi_1, \xi_2\right) = \mathrm{Cov}\,\left(\Phi^{-1}(U_1), \Phi^{-1}(U_2)\right),$$

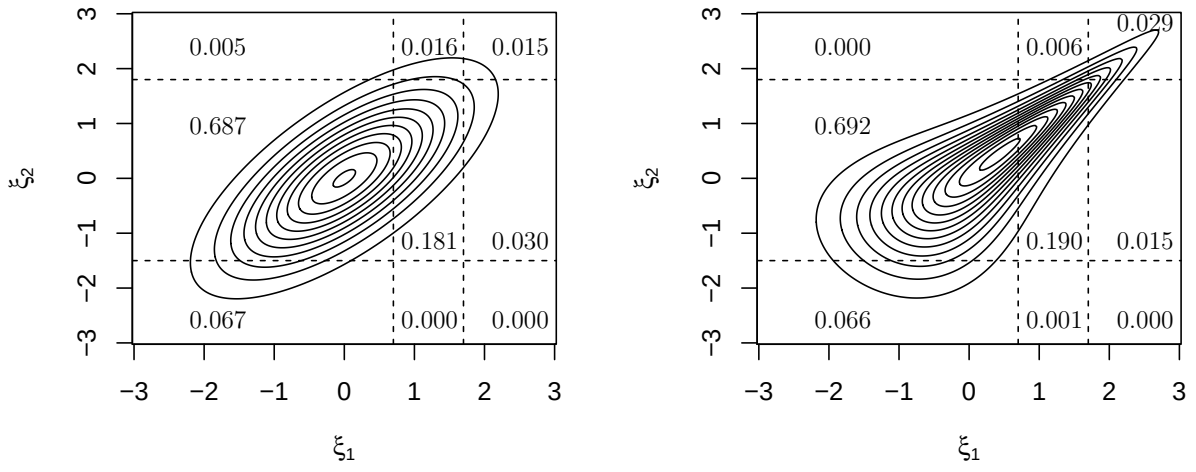
when (U_1, U_2) is distributed according to the Joe copula with dependence parameter $\theta = 3.011$.

Note that while the marginal distributions of (ξ_1, ξ_2) are exactly standard normal, its full bivariate distribution is far from normal. To visualize the difference, consider the contour lines of the density of H in eq. (2) presented in the right-hand panel of Figure 1. For comparison the contours of the bivariate normal distribution with standard normal marginals and correlation 0.7 are depicted in the left-hand panel. Clearly, although the two distributions have the same univariate marginals, and the same correlation of 0.7, the distributions are not the same. One notable feature of the Joe copula is the strong dependence in the upper tails.

The dashed horizontal and vertical lines in the figure represent the *thresholds* used in our example. These thresholds are cut-off values to discretize ξ_1 and ξ_2 into ordinal variables X_1 and X_2, respectively, each having three categories. As a special case of the upcoming general description (see eq. (3)), we have

$$X_1 = \begin{cases} 1, & \text{if } \xi_1 \leq 0.7 \\ 2, & \text{if } 0.7 < \xi_1 \leq 1.7\,, \\ 3, & \text{if } \xi_1 > 1.7 \end{cases} \qquad X_2 = \begin{cases} 1, & \text{if } \xi_2 \leq -1.5 \\ 2, & \text{if } -1.5 < \xi_2 \leq 1.8\,. \\ 3, & \text{if } \xi_2 > 1.8 \end{cases}$$

The resulting distribution of X_1 is given by the probabilities $0.758, 0.197, 0.045$, and is highly skewed. The distribution of X_2 is more symmetrical, with probabilities $0.067, 0.897, 0.036$. With the copula package (Hofert et al., 2013) we can calculate the joint probability distribution of X_1 and X_2, expressed as a 3×3 table. The values of the joint probabilities (rounded to three decimal points) are printed in the corresponding cells in Figure 1. We see that the joint probabilities of (X_1, X_2) when discretizing the normal



(a) Bivariate normal distribution

(b) Bivariate distribution with Joe copula

Figure 1. Contour lines for two bivariate distributions with correlation 0.7 and standard normal marginals. The dashed lines represent threshold values 0.7, 1.7 for ξ_1, and $-1.5, 1.8$ for ξ_2. In each cell defined by the thresholds is printed the corresponding joint probability for ordinal variables X_1 and X_2.

distribution differ somewhat from the probabilities obtained when discretizing the Joe distribution.

We now study the normal-theory (NT) likelihood function for observations of (X_1, X_2), as studied in Olsson (1979). This likelihood function is based on the assumption that (ξ_1, ξ_2) is bivariate normal with standardized marginals and Pearson correlation ρ, and is a function of hypothesized thresholds and Pearson correlation, and depends on the observations only through the 3×3 table of empirical proportions.

We note that the NT likelihood function is not a correctly specified likelihood function when (ξ_1, ξ_2) has a Joe copula. Therefore the NT maximum likelihood estimator (NT-MLE) does not need to be consistent for the actual Pearson correlation of (ξ_1, ξ_2), which we recall is 0.7 under both distributions, but will instead estimate a so-called least

false parameter with respect to the Kullback–Leibler divergence (see Jin & Yang-Wallentin (2017) for a discussion of this topic in the context of polychoric correlations and Claeskens & Hjort (2008) for a general perspective).

In order to compute the limits of the NT-MLE of ρ, we may approximate it by computing the NT-MLE from a very large simulated sample from (X_1, X_2). Alternatively, we may in this case identify this limit exactly through the following simple procedure. In the NT likelihood function, we insert the true 3×3 probability table, i.e. the population distribution of (X_1, X_2). The NT likelihood function can then be maximized, which results in the NT-MLE of an infinitely large sample of (X_1, X_2), i.e. the least false values in the population. We follow standard practice and use the two-stage method of Olsson (1979) and not the simultaneous MLE, and note that the least false parameter values of these two methods may differ (Jin & Yang-Wallentin, 2017).

We will call the limit of the NT-MLE of ρ for the NT polychoric correlation, and we note that this is a population parameter which need not equal the Pearson correlation of (ξ_1, ξ_2). Under the bivariate normal distribution shown in Figure 1(a), the NT polychoric correlation is, as dictated by theory, $\rho = 0.7$. Under the bivariate distribution with Joe copula in Figure 1(b) the NT polychoric correlation is instead $\rho = 0.99$. That is, the NT polychoric estimator has a substantive positive bias of 34%, which is close to the theoretical maximum, since 1 is the theoretical upper bound for correlations.

For the same distributions for (ξ_1, ξ_2) depicted in Figure 1 we might change the thresholds in order to obtain a substantial negative bias. With thresholds $-1.8, 1.8$ for ξ_1 and $-1.8, -1.2$ for ξ_2, the NT polychoric correlation is still 0.7 under the bivariate normal distribution, compared to 0.35 for the Joe distribution. The relative bias is now negative and very substantial at 50%.

A study of the impact of threshold configurations and underlying Pearson correlation on normal theory polychoric correlation bias

In the above we considered how the NT polychoric correlation behaves under two different ordinal distributions when ξ is non-normal. The only difference between the two distributions was the choice of thresholds, and the result was a substantial difference in NT polychoric correlation bias. Here, we report a more systematic exploration of the consequences of the placements of the thresholds, keeping the distribution of ξ fixed to a Joe copula with standard normal marginals and Pearson correlation of 0.7. We study how the bias in NT polychoric correlations vary across different threshold patterns. We also assess the effect of fixing the thresholds, and thereby also the distributions, to be equal for the two marginals.

Our approach is to randomly generate thresholds. For each threshold configuration we calculate the NT polychoric correlation for the resulting ordinal distribution. Note that while the thresholds will be randomly drawn in a manner described below, this is the only source of randomness. That is, we do not here assess the finite sample behavior of the polychoric estimator based on simulated data, but precisely calculate its limit.

For $K = 3$, only two actual thresholds are generated. In general, $K - 1$ thresholds are drawn. To illustrate our simulation method, we consider the $K = 3$ case in detail. Here, we draw random thresholds $\tau_{1,1}, \tau_{1,2}$ for marginal and $\tau_{2,1}, \tau_{2,2}$ for the second. The population value of the limit of the NT polychoric estimator is then calculated based on the distribution of the random variables

$$X_1 = \begin{cases} 1, & \text{if } \xi_1 \leq \tau_{1,1} \\ 2, & \text{if } \tau_{1,1} < \xi_1 \leq \tau_{1,2} \\ 3, & \text{if } \xi_1 > \tau_{1,2} \end{cases}, \qquad X_2 = \begin{cases} 1, & \text{if } \xi_2 \leq \tau_{2,1} \\ 2, & \text{if } \tau_{2,1} < \xi_2 \leq \tau_{2,2} \\ 3, & \text{if } \xi_2 > \tau_{2,2} \end{cases}$$

The distribution of (ξ_1, ξ_2) is still kept to the distribution of eq. (2), i.e. a distribution with standard normal marginals, and a Joe copula with $\theta = 3.011$. We also consider cases where $K = 5$ and $K = 15$. The simulation proceeds as above, except we use the general set up

described in eq. (3) in the upcoming section, generalizing the above equations to a finer discretization.

Thresholds are randomly generated in three ways, and this is summarized here. For more details, the R code in the supplementary material should be consulted. A first method for generating thresholds is by using uniform thresholds, which corresponds to simulating thresholds uniformly on $[-3, 3]$ and then sorting the numbers. The second method for generating thresholds is by using logarithmic thresholds. For $i = 1, 2$, we set $\tau_{i,k} = (\log k) - U_i$ where U_1, U_2 are independent and uniform on $[0, 3]$. The third method for generating thresholds produces symmetric thresholds. We consider cases where $K = 3, 5, 15$, and so $K - 1$ is an even number. The first $(K - 1)/2$ thresholds are generated by simulating random numbers uniformly and then sorting the numbers. The remaining thresholds are set equal to the first, but in reverse order, and with opposite signs. As a consequence, the distribution of the ordinal variables will also be symmetrical.

We also consider the condition where the marginals are equal by using the same thresholds. In this case we first generate $\tau_{1,1}, \tau_{1,2}$ and then set $\tau_{2,j} = \tau_{1,j}$ for $j = 1, 2$. In summary, for each level of K, by crossing equality/inequality of marginals with threshold generation approach, we have a total of six conditions. In each condition 1000 thresholds configurations were drawn, and the NT polychoric correlation was calculated for each configuration.

The resulting NT polychoric correlations are given in Figure 2. We first note that the symmetric case has less bias than the other cases. Also, equality of marginals reduces bias considerably in the symmetric case, but not in the asymmetrical cases. In the uniform and logarithmic conditions, the ranges of NT polychoric correlations are very wide, encompassing weak correlations and almost a perfect positive correlation. We finally observe that bias is reduced as the number of levels increases.

We repeated the study with the dependence parameter for the Joe copula set to $\theta = 1.53$, giving a Pearson correlation of 0.36 when paired with standard normal marginals.

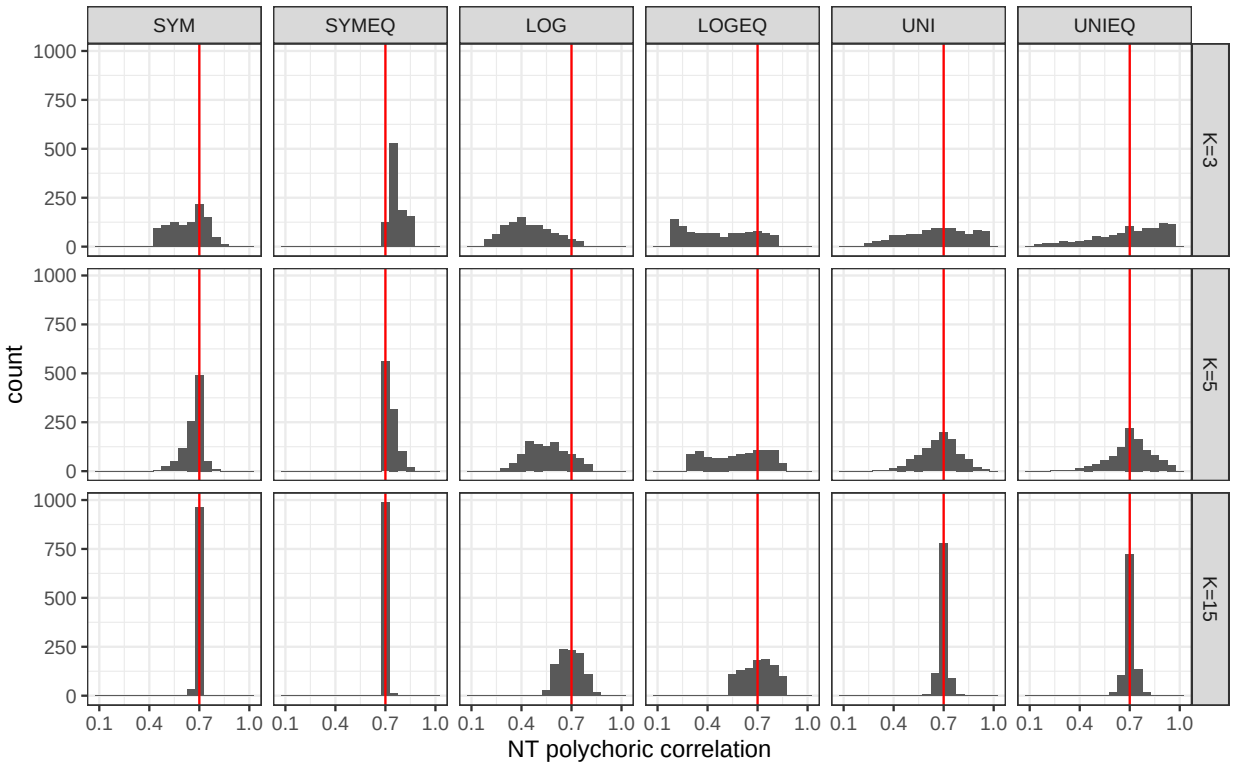


Figure 2. Histograms of NT polychoric correlations for underlying Joe distribution with correlation 0.7, represented as a red vertical line. SYM, LOG and UNI refer to symmetric, logarithmic and uniform threshold generation methods, respectively. The suffix -EQ refers to equal thresholds in the two marginals.

This more modest correlation leads to a distribution which is closer to standard Normal. Contour graphs of bivariate normal and Joe distributions for a Pearson correlation of 0.36 may be found in the online supplementary material (Figure S1), as well as histograms of the spread of NT polychoric correlations (Figure S2).

The findings in this low-correlation version of the experiment are similar to those in the above high correlation case. We observe that considerable variation in NT polychoric correlation also exists here across the randomly generated thresholds. When the number K of categories increase, the bias is generally reduced. However, the smaller bias associated with symmetric marginals in the high-correlation case is not observed in the low-correlation case. Also, equal marginal distributions in the symmetric case are not associated with

smaller bias, as was the case for the high-correlational case.

Finally, instead of varying the thresholds, let us consider them as fixed. By systematically varying the dependence parameter θ in the Joe copula, and calculating the NT polychoric correlation for each value of θ, we may study how it is affected by varying degrees of dependencies. We used fixed logarithmic thresholds, unequal in the two marginals, with the same random numbers U_1, U_2 for each $K = 3, 5, 15$. For a grid of $\theta \in [1, 18]$, we computed the resulting Pearson correlation for a Joe copula with dependence parameter θ and standard normal marginals, as well as the NT polychoric correlations for the given thresholds. The results are plotted in Figure 3. We see that the bias in general decreases with the number of thresholds, as is expected from an inspection of Figure 2. We also see that the bias is non-linear, being the greatest for Pearson correlations around 0.8, for this particular distributional configuration. For very high correlations, the bias decreases. In fact, for $\theta = 1$ and for $\theta = \infty$, the Joe distribution is a bivariate normal distribution: For $\theta = 1$ we have independence in the Joe copula, while for $\theta = \infty$ we have the so-called Frechet upper bound distribution with standard normal marginals (Joe, 1997, Chapter 3). Since the marginals are standard normal, this corresponds to a bivariate normal distribution with $\rho = 0$ and $\rho = 1$, respectively.

In conclusion, these examples show that estimating the polychoric correlation coefficient while assuming underlying normality can lead to substantive estimation bias (both negative and positive) when the normality assumption is violated. The wide range of possible NT polychoric correlations observed for $K \leq 5$ clearly demonstrates that the interaction between a specific threshold configuration and the underlying distribution has a strong effect on the bias of NT polychoric correlation. The nature of this interaction seems complicated, and we could only discern with clarity the following pattern: bias is consistently reduced when the number of categories increased. Also, symmetrical thresholds are associated with smaller bias, especially in the high-correlational case.

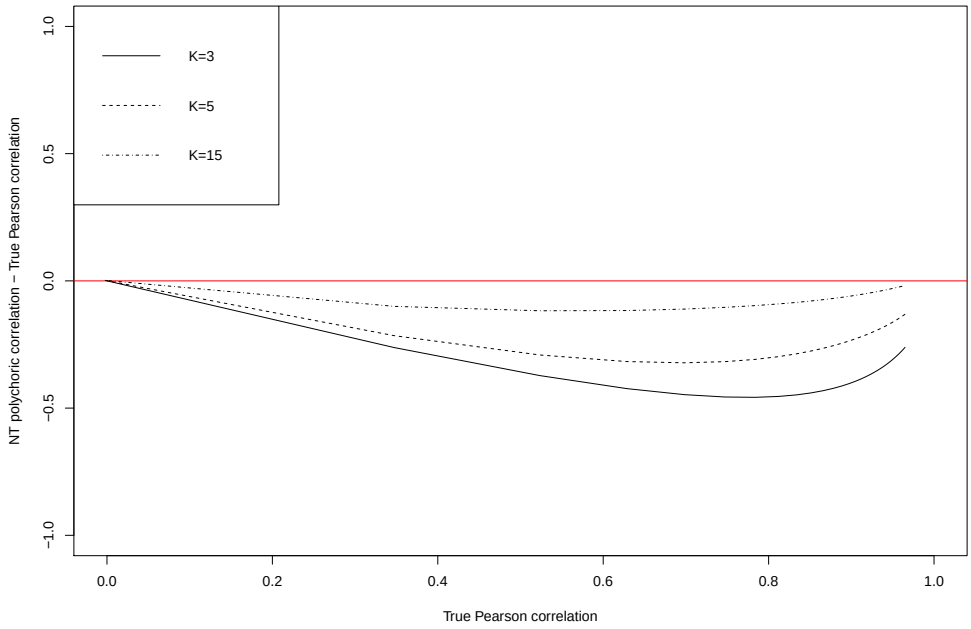


Figure 3. Variation of bias in NT polychoric correlation for a selection of Joe copulas with standard normal marginals and fixed logarithmic thresholds for $K = 3, 5, 15$.

Discretized normality and the test of Maydeu-Olivares

In ordinal SEM based on polychoric correlations the following framework is adapted. We observe n IID observations of a random vector $X = (X_1, X_2, \ldots, X_p)'$ whose coordinates are ordinal. We assume X is discretized from a continuous p-dimensional random vector ξ, which is hypothesized to be multivariate normal. This means that for $i = 1, 2, \ldots, p$, we have

$$X_i = \begin{cases} x_1, & \text{if } \tau_{i,0} < \xi_i \leq \tau_{i,1} \\ x_2, & \text{if } \tau_{i,1} < \xi_i \leq \tau_{i,2} \\ \vdots & \\ x_K, & \text{if } \tau_{i,K-1} < \xi_i \leq \tau_{i,K}. \end{cases} \tag{3}$$

We have for each i that $\tau_{i,0} = -\infty < \tau_{i,1} \leq \tau_{i,2} \leq \cdots \leq \tau_{i,K-1} \leq \tau_{i,K} = \infty$. Under this model the distribution of X is a function of the distribution of ξ and of the thresholds. By

identifiability considerations, we assume that ξ has standardized marginals. By the normality assumption, this means that the correlation matrix of ξ and the thresholds fully describe the distribution of X.

Maydeu-Olivares (2006) proposed a test for the null hypothesis that ξ has multivariate normal distribution. The test statistic is based on the discrepancy between the observed bivariate proportions in the sample and the probabilities implied by assuming that ξ is multivariate normally distributed. Let $k \neq l$ and denote by $p_{kl,ij}$ the number of observations in the sample with $X_k = i$ and $X_l = j$, divided by the sample size. These are the observed bivariate proportions in our dataset. To obtain the model-implied proportions, that is, the proportions we would expect under the assumption of discretized normality, we estimate the thresholds $\hat{\tau}$ and the polychoric correlation $\hat{\rho}$ between ξ_k and ξ_l (Olsson, 1979). Then, the model-implied proportion is calculated as $\pi_{kl,ij} = P(\hat{\tau}_{l,i-1} < \xi_l \leq \hat{\tau}_{l,i}, \hat{\tau}_{k,j-1} < \xi_k \leq \hat{\tau}_{k,j})$, assuming that (ξ_k, ξ_l) is normally distributed with covariance matrix $\left(\begin{smallmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{smallmatrix}\right)$. Note that in the probability defining $\pi_{kl,ij}$, threshold parameteres are estimated from data and are treated as fixed and their distribution is not included in the probability calculation. Let $r_{kl,ij} = p_{kl,ij} - \pi_{kl,ij}$ be the residual between the observed proportion and the proportion implied by normality. There are $p(p-1) \cdot K^2/2$ such residuals. Maydeu-Olivares (2006) derived the asymptotic distribution of the squared residuals when ξ is multivariate normal:

$$T := n \sum r_{kl,ij}^2 \quad \xrightarrow{d} \quad \sum_{i=1}^{d} \alpha_i Z_i^2, \quad \text{where } Z_1, Z_2, \ldots, Z_d \text{ IID } N(0,1), \tag{4}$$

where $d = (K^2 - 2K)p(p-1)/2$, and where $\alpha_1, \ldots, \alpha_d$ are the eigenvalues of the matrix

$$M = (I - \Delta G)\Gamma(I - \Delta G)', \tag{5}$$

where I is the identity matrix, and Δ is a Jacobian matrix defined as $\partial\pi/\partial\kappa$, where π contains the model-implied bivariate proportions, and κ contains the thresholds and the polychoric correlations. The matrix G is such that $\sqrt{n}(\hat{\kappa} - \kappa_0) \stackrel{a}{=} G\sqrt{n}(p - \pi_0)$, where π_0 contains the true bivariate proportions (Maydeu-Olivares, 2006, eq. 14). Since the limit in

eq. (4) is a mixture of independent chi-square distributions, Maydeu-Olivares (2006) proposed, in analogy with well-known approximations such as the Satorra-Bentler adjustment (Satorra & Bentler, 1988), to scale T so that its asymptotic mean and possibly variance equals that of a chi-square distribution with d degrees of freedom. A simple mean-scaling then yields

$$T_S = \frac{d \cdot T}{\mathrm{tr}(M)}.$$

A mean and variance correction obtained by scaling and shifting (Asparouhov & Muthén, 2010) yields

$$T_{SS} = c_1 \cdot T + c_2,$$

where

$$c_1 := \sqrt{\frac{d}{\mathrm{tr}(M^2)}}, \qquad c_2 := d - \sqrt{\frac{d \cdot \mathrm{tr}(M)^2}{\mathrm{tr}(M^2)}}.$$

We note that T_{SS}, although mean-and-variance adjusted, is different from the mean-and-variance adjusted test studied by Maydeu-Olivares (2006). However, these statistics have been empirically shown to be tightly related (Foldnes & Olsson, 2015).

To the best of our knowledge and aforementioned, no simulation study has evaluated the performance of these tests, with the exception of a small study ($p = 12$) in Maydeu-Olivares (2006), which used the multivariate t-distribution to evaluate the power of the tests. The results were promising, the tests maintained Type I error rates at an acceptable level, while exhibiting a high power to detect that the multivariate t-distribution was used, especially at the largest sample size. However, in that small study the number of categories was restricted to $K = 3$ and only symmetrical thresholds were considered.

A parametric bootstrap test for discretized normality

Motivation for using the parametric bootstrap

Given an ordinal dataset, let T^{obs} be the numerical value of the test statistic in eq. (4), and let $\tilde{T}^{\mathrm{obs}}$ denote the scaled statistic T_S or some other approximation to the true

distribution of T. The true p-value is $P(T > T^{\text{obs}})$, which may be approximated, using the large sample result of eq. (4), by

$$P(\chi^2_d > \tilde{T}^{\text{obs}}). \tag{6}$$

Since the limit variable in eq. (4) is not distributed according to a χ^2_d-distribution, except if e.g. $\alpha_1 = \cdots = \alpha_d$, the difference between the approximate and the actual p-value will not go to zero, i.e. we do not have consistency. Note that this potential inconsistency is a direct consequence of the chosen approximation, and that such approximations often work very well in finite samples (Foldnes & Grønneberg, 2017a).

A consistent method based on eq. (4) is obtained as follows. We estimate each $\alpha_1, \ldots, \alpha_d$ from data, resulting in estimates $\hat{\alpha}_1, \ldots, \hat{\alpha}_d$, and use the approximation

$$P(T > T^{\text{obs}}) \approx P\left(\sum_{i=1}^{d} \hat{\alpha}_i \chi^2_1 > T^{\text{obs}}\right),$$

where the probability treats $\hat{\alpha}_1, \ldots, \hat{\alpha}_d$ as fixed. This test, in the context of model fit for structural equation models, was studied by Foldnes & Grønneberg (2017a). The problem with this approach is that $d = (K^2 - 2K)[p(p-1)/2]$, a high number even in moderately simple problems. It therefore seems likely that the estimation of $\alpha_1, \ldots, \alpha_d$ will introduce a large degree of approximation error compared to using eq. (6).

Description of the proposed test

We next propose an alternative and consistent approximation to the p-value based on the parametric bootstrap. Under the null hypothesis of discretized normality, the distribution of the data is fully specified by a parametric model, and our proposed test is therefore a simple application of the parametric bootstrap, see e.g. Efron & Tibshirani (1994) and Rice (2006). Note that since we can simulate from the parametric model using the estimated parameters, we do not need to re-sample from the data. Such re-sampling is connected with applying the bootstrap to non-parametric models (Efron & Tibshirani, 1994). We now give a high level overview of the proposed method. A detailed technical description of the parametric bootstrap test is given in Algorithm 1.

Under the null hypothesis that ξ is normally distributed with standardized marginals, the underlying probability distribution of the data is of the form

$$P_{\Sigma,\tau}(\cdot)$$

where Σ is the correlation matrix of ξ and τ is the vector of the thresholds for the representation of X in eq. (3). The true, unknown p-value is therefore

$$p_{\text{true}} = P(T > T^{\text{obs}}) = P_{\Sigma,\tau}(T > T^{\text{obs}}).$$

The parametric bootstrap p-value uses estimators as approximations for Σ and τ. For concreteness, assume that $\hat{\Sigma}, \hat{\tau}$ are obtained with the normality-based method of Olsson (1979). The bootstrap p-value is

$$p_{\text{boot}} = P_{\hat{\Sigma},\hat{\tau}}(T > T^{\text{obs}}), \tag{7}$$

and is consistent. Since the probability is not easily calculated exactly, we follow standard bootstrap practice (Efron & Tibshirani, 1994) by simulating B (where B is an appropriately high number) new samples of size n, generated by discretizing a multivariate normal vector with standardized marginals and correlation matrix $\hat{\Sigma}$. The discretization is done using eq. (3) with thresholds from $\hat{\tau}$. We compute the test statistics $T_1, T_2, \ldots, T_B$, each being computed on the basis of the simulated samples. Using these B simulated test statistics, we approximate the probability in eq. (7) using

$$\hat{p}_{\text{boot}} = \frac{1}{B} \sum_{i=1}^{B} I\{T_i > T^{\text{obs}}\}$$

where $I\{A\}$ is the indicator function of A, which is 1 if A is true, and zero otherwise. By the law of large numbers, $\hat{p}_{\text{boot}}$ is a consistent approximation of p_{boot} as $B \to \infty$, which in turn approximates p_{true}. The stepwise progression of the test is outlined in Algorithm 1, terminating with the p-value associated with the null hypothesis of underlying normality.

Simulating non-normal variables for ordinal covariance modeling

To evaluate the test of Maydeu-Olivares (2006) and the bootstrap test we apply Monte Carlo simulation. We simulate non-normal continuous data with a pre-specified

Algorithm 1 Bootstrap test of underlying normality

procedure BOOT(original sample)
  Calculate the polychoric correlation matrix $\hat{\Sigma}$, and the threshold sets $\hat{\tau}_i$ for each variable, based on original sample
  Use original sample to calculate T_{orig} with eq. (4)
  for $k \leftarrow 1, \ldots, B$ **do**
    cont.sample $\leftarrow$ A random sample drawn from $\text{N}(0, \hat{\Sigma})$
    ordinal.sample $\leftarrow$ Discretize cont.sample using thresholds $\hat{\tau}_i$
    Use ordinal.sample to calculate T_k with eq. (4)
    $\text{P}(k) \leftarrow 1$ if $T_k > T_{\text{orig}}$, 0 otherwise
  end for
  return the p-value as $\frac{\sum_{k=1}^{B} P(k)}{B}$
end procedure

correlation matrix, which is subsequently discretized using various threshold configurations. Our goal is to control the univariate ordinal distributions, and the underlying correlation matrix, while increasing the degree of non-normality in ξ. In this way the correlational structure of ξ and the observed ordinal distributions are kept constant, and will not be confounded with the effect of non-normality in ξ. Following identifiability arguments given in Foldnes & Grønneberg (2019), we restrict attention to distributions of ξ that have standard normal marginals.

There are many options for distributions with a prespecified correlation matrix Σ and with standard normal marginals. The multivariate normal distribution with covariance matrix equal to Σ and with expectation equal to the p-dimensional zero vector is one candidate. A random vector with this distribution is denoted by Z in the following. To break the underlying normality assumption, but still respect that the univariate marginals be $N(0, 1)$ and the correlation matrix be Σ, we used the VITA (VIne To Anything) method (Grønneberg & Foldnes, 2017). VITA distributions are so-called regular vines (Bedford et

al., 2002) with a pre-specified covariance matrix. Regular vines are, briefly summarized, multivariate distributions constructed from bivariate copulas according to a hierarchical scheme that may be visualized as a sequence of tree graphs. The nodes in the first tree represent variables. In the next tree, the nodes represent unconditional pairs of variables. In subsequent trees the nodes represent pairs of variables that are conditional on sets of variables. An illustration of such a hierarchy of trees in the five-dimensional case is given in Figure 4. This hierarchy of trees will be used in our Monte Carlo study. In addition to the hierarchical tree structure, the user specifies univariate marginal distributions and the type of bivariate copulas to be used for each edge in the trees. A numerical search across the trees is performed to calibrate the copula parameters so that a desired correlation matrix is reached. As argued by Foldnes & Grønneberg (2019), the VITA method is especially well suited for the problem at hand, as the marginals can be fixed to standard normal, and we have specified a correlation matrix. What is left is to specify a sequence of trees, and a family of bivariate copulas to use for constructing the VITA distribution.

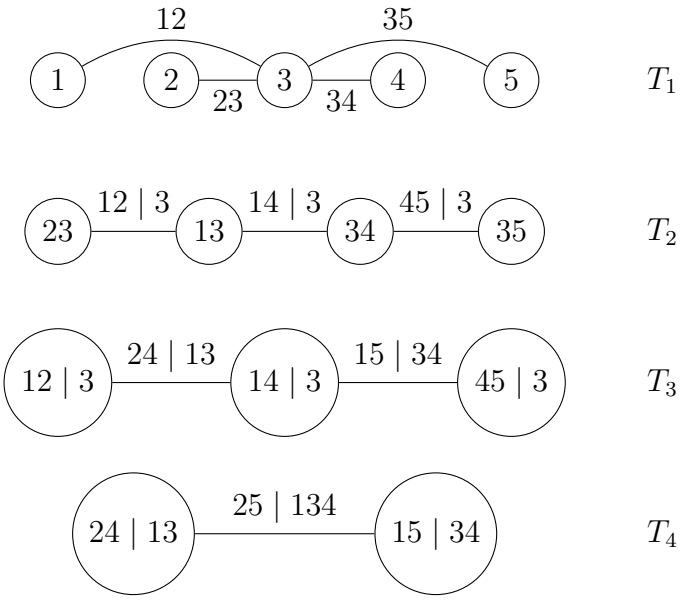


Figure 4. A five-dimensional regular vine.

Simulation design

Next we present the design conditions for our Monte Carlo study.

Dimensionality and correlation matrix of ξ

We generated data with dimensions $p = 5$, $p = 10$ and $p = 15$. These dimensionalities represent a range from small- to medium-sized models used in CFA studies. Given the considerable computational running time of the tests evaluated in this study when dimensionality increases, we did not include sample sizes that reflect larger ($p \geq 20$) models. Test evaluation at larger dimensions is a worthy topic for future studies. In all conditions the univariate marginals were of standard normal distribution, and at each dimension p, the correlation matrix of ξ was kept fixed. The correlation matrices were obtained as model-implied matrices for two-factor analytical models; see the appendix for R code. For the smallest case the correlation matrix Σ_5 of the discretized variable $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)'$ is given in Table 1. For dimensions $p = 10$ and $p = 15$, the correlation

	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
ξ_1	1.00				
ξ_2	0.56	1.00			
ξ_3	0.48	0.42	1.00		
ξ_4	0.40	0.35	0.30	1.00	
ξ_5	0.32	0.28	0.24	0.20	1.00

Table 1

Correlation matrix Σ_5 of ξ for the $p = 5$ case.

matrices Σ_{10} and Σ_{15} are given in Tables 2 and 3, respectively. We remark that all correlations are weak to moderate.

	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	ξ_9	ξ_{10}
ξ_1	1.00									
ξ_2	0.56	1.00								
ξ_3	0.48	0.42	1.00							
ξ_4	0.40	0.35	0.30	1.00						
ξ_5	0.32	0.28	0.24	0.20	1.00					
ξ_6	0.19	0.17	0.14	0.12	0.10	1.00				
ξ_7	0.17	0.15	0.13	0.10	0.08	0.56	1.00			
ξ_8	0.14	0.13	0.11	0.09	0.07	0.48	0.42	1.00		
ξ_9	0.12	0.10	0.09	0.07	0.06	0.40	0.35	0.30	1.00	
ξ_{10}	0.10	0.08	0.07	0.06	0.05	0.32	0.28	0.24	0.20	1.00

Table 2

Correlation matrix Σ_{10} *of* ξ *for the* $p = 10$ *case.*

Distribution of ξ

Fixing the marginals to standard normal and fixing the correlation matrix still allow for many feasible distributions for ξ. For each level p of dimensionality we generated data from five such distributions. To study Type I error control of the tests, the first distribution for ξ was the multivariate normal. To allow for power investigations, four non-normal distributions were considered. These were based on VITA distributions. We restricted ourselves to two types of bivariate copulas when constructing vines: the Gumbel family of copulas and the Joe family of copulas, see Nelsen (2007) and Grønneberg & Foldnes (2017) for technical definitions. These copulas differ from the normal copula by, e.g. allowing tail dependencies. Let us for $p = 5, 10$ and 15 denote by Z_p a multivariate normal p-vector with standard normal marginals and whose correlation matrix equals Σ_p. Likewise let G_p denote a random p-vector whose distribution equals that of the regular vine constructed from Gumbel pair-copulas, and let J_p denote a random p-vector whose

	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	ξ_9	ξ_{10}	ξ_{11}	ξ_{12}	ξ_{13}	ξ_{14}	ξ_{15}
ξ_1	1.00														
ξ_2	0.56	1.00													
ξ_3	0.48	0.42	1.00												
ξ_4	0.40	0.35	0.30	1.00											
ξ_5	0.32	0.28	0.24	0.20	1.00										
ξ_6	0.19	0.17	0.14	0.12	0.10	1.00									
ξ_7	0.17	0.15	0.13	0.10	0.08	0.56	1.00								
ξ_8	0.14	0.13	0.11	0.09	0.07	0.48	0.42	1.00							
ξ_9	0.12	0.10	0.09	0.07	0.06	0.40	0.35	0.30	1.00						
ξ_{10}	0.10	0.08	0.07	0.06	0.05	0.32	0.28	0.24	0.20	1.00					
ξ_{11}	0.19	0.17	0.14	0.12	0.10	0.19	0.17	0.14	0.12	0.10	1.00				
ξ_{12}	0.17	0.15	0.13	0.10	0.08	0.17	0.15	0.13	0.10	0.08	0.56	1.00			
ξ_{13}	0.14	0.13	0.11	0.09	0.07	0.14	0.13	0.11	0.09	0.07	0.48	0.42	1.00		
ξ_{14}	0.12	0.10	0.09	0.07	0.06	0.12	0.10	0.09	0.07	0.06	0.40	0.35	0.30	1.00	
ξ_{15}	0.10	0.08	0.07	0.06	0.05	0.10	0.08	0.07	0.06	0.05	0.32	0.28	0.24	0.20	1.00

Table 3

Correlation matrix Σ_{15} *of* ξ *for the* $p = 15$ *case.*

distribution equals that of the regular vine constructed from Joe pair-copulas. Note that Z_p, G_p and J_p all have standard normal marginals and correlation matrix Σ_p. For technical details in the construction of these distributions, see the R code in the supplementary material. In order to obtain a distribution whose non-normality is intermediate between Z_p and G_p we also simulated data from the random vector

$$ZG_p = \frac{1}{\sqrt{2}} Z_p + \frac{1}{\sqrt{2}} G_p,$$

whose marginal distributions are standard normal and whose correlation matrix equals Σ_p, by construction. Likewise, to interpolate between normality and the distribution of J_p we

simulated from

$$ZJ_p = \frac{1}{\sqrt{2}}Z_p + \frac{1}{\sqrt{2}}J_p.$$

To sum up, for $p = 5, 10$ and 15, we simulated from the following random vectors: Z_p, ZG_p, G_p, ZJ_p and J_p. By allowing two kinds of non-normal distributions, in the form of G_p and J_p, we may investigate how different types of non-normality affect the performance of the test of Maydeu-Olivares (2006) and the bootstrap test. Also, by allowing an intermediate distribution between the normal case Z_p and each of G_p and J_p, we may better study the power of the tests to detect non-normality as we progressively move from Z_p to ZG_p and finally to G_p, and likewise from Z_p to ZJ_p and finally to J_p.

Number of levels and distributions of observed variables

In empirical research the most common numbers of levels in ordinal data are $K = 4, 5$ and 7 (Li, 2016b). We included these levels in our study. For each K, we considered three configurations of thresholds, resulting in three distributions for the ordinal marginals in X, which we refer to as symmetrical, moderately skewed and strongly skewed. The corresponding distributions of the discretized variables are presented in Figure 5.

In accordance with earlier studies (e.g. Flora & Curran, 2004; Li, 2016a,b) we keep all marginals fixed. Considering the numerical experiments reported in Figure 2, it appears that we do not lose much generality by not considering unequal thresholds across marginals in the skewed conditions. However, in the symmetrical conditions equal marginals may yield smaller bias.

Sample sizes

Sample size will affect test performance. The larger the sample, the larger the power to detect underlying non-normality is expected to be. In the current study we included two sample sizes, 200 and 1000, which we regard as representing a relatively small and a relatively large sample size, respectively.

Data generation and test implementation

For the smallest dimension, $p = 5$, we included a full factorial design, with 5 (distributions for ξ) $\times$ 3 (number of levels in observed variables) $\times$ 3 (ordinal observed distributions) $\times$ 2 (sample size) = 90 experimental conditions. At the larger dimensions we excluded the $K = 5$ condition and in addition, for $K = 4$ and $K = 7$ we excluded the moderately skewed ordinal distribution. Hence, for each of $p = 10$ and $p = 15$, there were 5 (distributions for ξ) $\times$ 2 (number of levels in observed variables) $\times$ 2 (ordinal observed distributions) $\times$ 2 (sample size) = 40 experimental conditions. The reason for considering a fewer number of conditions for $p = 10$ and $p = 15$ was related to computational resource restrictions in performing the simulation studies. Both the test of Maydeu-Olivares (2006) and the bootstrap test are time-consuming at high dimensions. In order to calculate the adjusted test statistics of Maydeu-Olivares (2006), large-dimensional matrix multiplications are needed in order to obtain M in equation (5). As an illustration, when $p = 15$ and $K = 7$, M has dimension of 5145×5145. The bootstrap test is also time-consuming, since it computes T in many bootstrap samples. However, the bootstrap procedure does not need to compute the matrix M. In addition the bootstrap procedure is easy to implement using parallel computing, which will reduce its running time considerably.

In total, 90+40+40=170 experimental conditions were considered. In each condition, 1000 random samples were drawn, and the outcomes of the tests of underlying normality were recorded. For the bootstrap test $B = 1000$ bootstrap samples were drawn for each generated sample.

Data generation and calculation of test statistics were performed in R using packages sirt (Robitzsch, 2019) and VineCopula (Schepsmeier et al., 2018), for polychoric estimates and for constructing and simulating from regular vines, respectively. The simulations were performed on the Abel Cluster, owned by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR)

Evaluation criterion

The performance of the tests was assessed in terms of rejection rates at the 5% level of significance.

Bias of polychoric correlations and standard errors across simulation conditions

Before we present simulation results on tests of underlying non-normality in the next section, we first investigate how our simulation conditions affect central ingredients in ordinal SEM, namely the polychoric correlations and their standard errors. From the practical perspective of ordinal SEM, it is important that conditions with substantial bias in the polychoric estimates and/or in their associated standard errors are detected by a test for underlying non-normality. On the other hand, if the bias is not substantial, we can accept from the practical perspective of ordinal SEM that underlying non-normality is detected less often. In conditions where we have substantive estimation bias with respect to the polychoric correlations, it is likely that also ordinal SEM will be biased. And from a practical point of view, it is these conditions that we hope the tests for non-normality will be able to detect.

Inference for the polychorics, and subsequently for ordinal SEM, may be incorrect in two ways. Underlying non-normality may induce substantive bias in the polychoric correlations, and in the standard error associated with a polychoric estimate. If the polychoric estimates become biased due to non-normality, SEM parameters are also likely to be biased and the ordinal SEM procedure is invalid. However, even with unbiased polychoric estimates, the inference based on these may be invalid due to non-normality. The reason is that parameter standard error and also the chi-square test of correct model specification are calculated from formulas where an estimate $\hat{\Gamma}$ of the asymptotic covariance matrix Γ of the polychoric estimator is plugged in (see Muthén (1978, 1984), as well as the general estimation and inference framework for covariance models found in Satorra (1989)). This matrix is estimated while assuming underlying normality. In conditions where $\hat{\Gamma}$ is

not a consistent estimator of Γ, due to underlying non-normality, the standard errors and the chi-square test statistic will not be reliable, and ordinal SEM inference will be invalid.

Bias in polychoric estimates

Figure 6 shows plots for dimension $p = 10$ of the difference between the estimated polychoric correlations and the true correlations of the underlying vector. For dimensions $p = 5$ and $p = 15$ the observed patterns are similar, with corresponding figures available in the supplementary material (Figures S3 and S4). In each combination of underlying distribution (Z, ZG, ZJ, G and J), number of categories ($K = 4, 5, 7$) and observed ordinal distribution (symmetrical, moderately and severely skewed), the estimated polychoric correlations were calculated from a large sample ($n = 10^6$).

It is seen that the polychoric estimates are not substantively biased under symmetrical and moderately skewed ordinal distributions, for all underlying distributions and observed ordinal distributions. However, under the severely skewed condition, non-normality implies negatively-biased polychoric correlations. This is a case that is often encountered in practice. As expected, the bias increases when we move from a vector combined from a normal and a VITA vector (ZG and ZJ) to the full VITA vector (G and J). Also, we note that the type of underlying normality (G vs. J) affects the bias. Under the Joe VITA vector the bias is more pronounced than under the Gumbel VITA vector, and this is mirrored when comparing ZG with ZJ. Across all conditions, the number of observed categories does not affect the bias in polychoric estimates.

Bias in standard error estimation

In all simulation conditions the true covariance matrix Γ of the polychoric estimates was compared to the normal-theory matrix Γ_{NT}. In statistical software the latter matrix, which assumes underlying normality, is estimated from data and used as an estimate of Γ. Discrepancies between the elements in Γ_{NT} and Γ imply that standard errors in ordinal SEM, being based on the assumption that $\Gamma = \Gamma_{NT}$, are invalid. In each simulation

condition we estimated Γ by simulating 20 000 samples, each of size $n = 4\ 000$. In each sample the vector of polychoric estimates was obtained, and we approximated Γ by the empirical covariance matrix of these vectors. A high-precision estimate of Γ_{NT} was obtained from a large sample ($n = 10^5$) using standard routines from lavaan.

Following the approach suggested by Foldnes & Grønneberg (2017b), the discrepancy between the matrices was visualized by plotting the difference between corresponding elements in Γ_{NT} and Γ against the elements of Γ. For dimension $p = 10$ this can be seen in Figure 7, where we have 45 polychoric correlations, and 1 035 non-redundant elements in Γ. Similar figures for $p = 5$ and $p = 15$ are included in the online supplementary material, and show the same overall pattern (Figures S5 and S6). As expected, with data drawn from the multivariate normal vector Z there is no systematic discrepancy between Γ_{NT} and Γ. It was found that bias in the normal-theory estimator of Γ was most pronounced under the symmetrical condition. This was unexpected given the unbiasedness of the polychoric estimator in this condition, see Figure 6. Under symmetrical ordinal distributions, for all values of K, there is substantive negative bias for many elements of Γ, especially under the full VITA vectors G and J. As we progress from symmetrical (SYM) to moderately and severely skewed distributions (SKEW1 and SKEW2, respectively), the bias progressively disappears. This trend is opposite to what we observed for the polychoric estimates in Figure 6, where the bias increased. In the condition of severe skewness, where the polychoric correlations were substantively biased under the full VITA vectors G and J, there is little bias in the normal-theory estimate of Γ. As was the case for polychoric correlation bias, the Joe VITA implies a larger bias than the Gumbel VITA. Also, the intermediate vectors ZG and ZJ generate less bias than the full VITA vectors G and J.

Relation between bias and type and degree of underlying non-normality

We have seen that in our simulation design, the degree of bias in the polychoric correlation coefficients and its normal theory covariance matrix Γ_{NT} is associated with the

degree of non-normality. For instance, moving from the multivariate normal condition Z via ZJ to the full VITA condition J entails a steady increase in bias. Also, whether the non-normal distribution is based on Gumbel or Joe pair-copulas affects the degree of bias. The bias summarized above is of a sufficient magnitude for it to be practically important, but is not dramatic compared to the introductory bivariate examples, see Figure 2. As illustrated in the bivariate example, it is easy to find simulation set ups where the bias is much more pronounced. Even when the thresholds are fixed, Figure 3 indicates that we here may dramatically increase the bias by using an example with higher correlations. The correlations in our simulation design (Tables 1, 2 and 3) are all small or moderate. We also did not search for thresholds which induce specifically severe biases, but used thresholds similar to those used in earlier papers (e.g Rhemtulla et al., 2012; Flora & Curran, 2004; Li, 2016b).

As we have seen in the bivariate illustrations, there is considerable variation in the bias of NT polychoric correlations, and presumably also of Γ_{NT}, originating from the placements of the thresholds. It is important to observe that this variability does not reflect a change in the degree of underlying non-normality, which is fixed by the distribution of ξ. The relationship between the bias and the degree of underlying non-normality is therefore complex, and it seems implausible that tests for underlying normality will always succeed in identifying cases with practically important bias in polychoric correlations and Γ_{NT} but will not reject for non-normal cases that do not induce such bias. What relation there is between the considered tests for underlying normality and variability of bias when varying the thresholds is investigated only to a small degree in the present study, and should be further investigated in a follow up study.

Simulation results: Type I control and power for testing underlying normality

We first report Type I error control of the tests, before reporting on power to detect underlying non-normality.

Type I error control

A commonly used criterion for acceptable Type I error control was proposed by Bradley (1978): the empirical rejection rate at 5% significance level should be between 2.5% and 7.5%. In Tables 4 and 5 we use gray backgrounds in cells that correspond to acceptable Type I error control.

For the lowest dimensionality, $p = 5$, empirical rejection rates at the 5% level of significance are given in Table 4. As was observed by Foldnes & Olsson (2015), the mean-adjustment in T_S leads to higher rejection rates than the mean-and-variance adjustment in T_{SS}. In fact, the rejection rate of T_S is unacceptably high under all conditions. In contrast, T_{SS} and the bootstrap test performs well in all conditions. It is noteworthy that across different numbers of categories and ordinal distributions, these two tests are highly successful in controlling the rate of Type I errors.

For higher dimensionalities, $p = 10$ and $p = 15$, the rejection rates are given in Table 5. As was the case for $p = 5$, T_S has a strong tendency to overreject the underlying normality hypothesis, while the tendency for T_{SS} is to underreject, as reported by Foldnes & Olsson (2015). The bootstrap test performs better than T_S and T_{SS}, exhibiting acceptable Type I error control in most conditions. However, at the smallest sample size the bootstrap test tends to overreject the null hypothesis. Across all conditions in Table 5, the poorest bootstrap test rejection rate is 0.086, which still is relatively close to the range of acceptable performance. In comparison, T_S and T_{SS} exhibit inacceptably high and low rejection rates, respectively, in almost all conditions. We also note that whether the ordinal distribution is symmetrical or not does not affect the rejection rates of the bootstrap and T_{SS}.

Power

We next report on the power of the tests to detect discretized non-normality. Due to the poor Type I error control of T_S we deem this test to be inadequate and do not report

its power.

Table 6 presents rejection rates at the 5% level of signficance for the lowest dimensionality. In the 72 conditions obtained by fully crossing 4 kinds of distributions, 2 sample sizes, 3 levels of categories and 3 types of thresholds the power to reject the underlying normality assumption is most often higher for the bootstrap test compared to the T_{SS} test. In fact, in only 9 of the 72 conditions did T_{SS} exhibit higher power than the bootstrap test. The generally higher rejection rates under the bootstrap mirror the findings observed under discretized normality, where we found that Type I error rates of the bootstrap were generally higher than those of T_{SS}.

As expected, the power of both tests increased when the distribution changes from ZG to G, and likewise from ZJ to J. That is, it is harder to detect non-normality of random vectors that are sums of a normal component and a non-normal component compared to random vectors composed entirely of the non-normal component.

We finally remark that the type of non-normality affects the power of both tests. The type of non-normality in the Joe VITA distribution is easier to detect than the non-normality in the Gumbel VITA distribution. For instance, when $n = 200$ and there are 7 symmetrical levels in the observed ordinal distribution, the bootstrap test only rejects 24% of the generated samples under the Gumbel VITA, compared to 76% under the Joe VITA.

The power results for the two largest dimensionalities are given in Table 7. Again, the bootstrap generally has higher a power to detect non-normality than T_{SS}. In all of the 64 conditions in Table 7 the bootstrap has power that is higher than or equal to that of T_{SS}. As expected, power increases with increasing sample size, and also when introducing more non-nonormality, i.e, when comparing J with ZJ, and G with ZG. Further, we observe that the non-normality of G is less likely to be detected compared to the non-normality of J. This is especially evident when $p = 15$ and $n = 200$. Particularly bad is the performance of T_{SS} in this condition under the Gumbel VITA distribution, never rejecting the null

hypothesis of underlying normality.

Power tends to decrease under asymmetrical compared to symmetrical ordinal distributions, at the smallest sample size. For instance, under $p = 10$, $n = 200$ and with $K = 4$, for the Joe VITA distribution the bootstrap rejection rates are 88% and 29%, under symmetrical and asymmetrical distributions, respectively. At the largest sample size this tendency is repeated when $K = 4$, but for $K = 7$ there are conditions where asymmetrical distributions lead to higher power compared to symmetrical distributions, e.g. under the ZJ distribution, $p = 15$ and $n = 1000$, the bootstrap test has a power of 69% under asymmetrical marginals compared to 45% under symmetrical observed marginals.

Finally, we note that power is generally higher with $K = 7$ categories compared to $K = 4$ categories, and especially for asymmetrical distributions.

Discussion

The non-normal simulation conditions employed in our simulation study have been shown to affect either the precision of the polychoric estimates, or the precision of their estimated standard errors. Especially surprising was the large number of conditions where the polychoric correlations were approximately unbiased, but where their associated standard errors were biased. However, this may be an artifact of our specific simulation design. Further research is needed to probe the complex interaction among observed distributional forms and type of underlying non-normality with respect to bias in polychoric inference. In general, non-normality was consistently observed to cause problems in inference relating to the polychoric correlations. Given the centrality of polychoric inference for ordinal SEM, we may conclude that ordinal SEM, within the context of our simulation design, in general is sensitive to underlying non-normality. This sensitivity depends on the type and degree of underlying non-normality, and – as a complicating factor – the placements of the thresholds, a factor which is completely unconnected to the degree of non-normality of ξ. We have observed that a VITA vector

constructed from Joe pair-copulas embeds a more challenging type of non-normality than a VITA vector constructed from Gumbel pair-copulas. Also, we have seen that an intermediate distributional type between a normal and a VITA vector causes less bias in polychoric inference than the full VITA vector.

To the best of our knowledge, we have conducted the first comprehensive evaluation of test statistics proposed by Maydeu-Olivares (2006) for underlying normality in ordinal data. We have also proposed an alternative in a new bootstrap test. Using a newly developed simulation methodology, we evaluated the tests in a Monte Carlo study where the underlying continuous distribution, the number of categories and the distribution of the ordinal marginals were manipulated. Two approximations to the test statistic of Maydeu-Olivares (2006) were evaluated. The mean-scaled approximation performed poorly, not being able to maintain type I error control. The scaled-and-shifted approximation performed better, and maintained type I error control in five dimensions. However, in dimensions 10 and 15 the scaled-and-shifted test exhibited poor Type I error control. In contrast, the new bootstrap test was able to control Type I error rates in almost all conditions. The bootstrap test also exhibited higher power than the scaled-and-shifted test to detect underlying non-normality. In summary, in the conditions employed in the present study, we found the bootstrap test to perform better overall than the large sample approximations to the distribution of the test statistic of Maydeu-Olivares (2006).

From the practical perspective of the researcher estimating an ordinal SEM to gain substantial knowledge in their field of study, an important question is whether a test of underlying non-normality can discern between conditions where non-normality is only mildly violated and ordinal SEM is approximately valid, and conditions where the degree and type of non-normality invalidate polychoric inference and thereby also ordinal SEM. Our proposed bootstrap test is based solely on statistical considerations, and ideally should therefore attain the highest possible power in all non-normal conditions. However, a test that detects even small deviations from normality will from the practical perspective be too

conservative. We observed that the bootstrap test has lower power under ZG compared to G, and under ZJ compared to J, so the test is indeed less strict when polychoric-related bias is less severe. Also, the test has lower power under ZG and G, compared to ZJ and J, respectively. Hence, the bootstrap test detects with higher probability the type of non-normality that produce the larger bias. Despite these findings, there are some observations that we did not expect and that we cannot currently explain. For instance, the power of the bootstrap test generally was highest under the moderately skewed condition in Table 7. Under this kind of observed ordinal distribution the bias in polychoric estimates (see Figure 6) and the bias in polychoric standard errors (see Figure 7) were moderate. That is, conditions where there is moderate bias in both polychoric correlations and standard errors are more likely to be detected by the bootstrap test than conditions with substantive bias in only polychoric correlations, or only in standard errors. The number of observed categories (4, 5 or 7) in general did not affect the polychoric inference or the performance of the bootstrap test. In contrast, the shape of the observed ordinal distribution was found to impact these outcomes substantially.

Generally, at the smallest sample size ($n = 200$) the bootstrap test had in many conditions low power to detect underlying non-normality, especially when the underlying distribution was half-way between normality and VITA non-normality. It is no surprise that the problem of detecting underlying non-normality in multivariate ordinal data is difficult, since so much of the structure in the continuous vector is lost when it is discretized. Our bootstrap test was based on bivariate probabilities. It will be a challenge for future research to develop a more powerful test, possibly by including trivariate probabilities in the test statistic.

Limitations

Although we simulated ordinal data for covariance modeling using the best available methods, we only evaluated two kinds of non-normality in the underlying discretized

vector. There are of course many possible non-normal distributions that can be constructed and simulated from using the VITA framework, and even more distributions outside the regular vines available with VITA. It is still not known which types of distributions are most often encountered in practice, and it is therefore difficult to know if our discretized VITA distributions may approximate real-world ordinal data. Another limitation is that only two sample sizes were considered. Also, we considered only discretized vectors whose pairwise correlations were quite moderate. In Tables 1 - 3 no two pairs of variables had a correlation higher than 0.56. In the case where the underlying ξ has pairs of variables with higher correlations, at least up to a certain point, the non-normality should be easier to detect than was the case in the present study, as shown in the bivariate illustration, see Figure 3 (p.15). A future Monte Carlo study may include correlational strength as an experimental factor to investigate this issue. Our study, in line with many previous studies, did not mix the number of categories or distributional shapes in a given simulation condition. As different variables in most practical settings have different distributional forms, future studies should systematically investigate whether mixing different ordinal distributions will aggravate the impact of non-normality on ordinal SEM beyond that observed in the present study.

An important limitation of the present investigation is that we did not consider cases with missing data. An overview of available methods is given in Jia & Wu (2019). We conjecture that such methods may be combined with a variant of the bootstrap methodology proposed in the present paper. A systematic formalization and evaluation of such an approach should be made in future studies.

Conclusion

Structural equation modeling with ordinal data is regularly based on an assumption of discretized normality. Due partly to the lack of statistical tests in currently available software, and to recommendations in previous simulation studies that suggest that ordinal

SEM is quite robust to underlying non-normality, the normality assumption has often been taken for granted by practitioners.However, recent research has unveiled problems with the simulation methodology in previous simulation studies. In the present study we adopted a recently proposed simulation method for ordinal data to study both the effect of underlying non-normality on polychoric estimation, and the performance of two tests for underlying normality. We found that polychoric correlation was very sensitive towards non-normality in ξ, and that the sensitivity was strongly moderated by the placements of the thresholds, i.e. the observed ordinal distributions.

The first test for underlying non-normality was proposed by Maydeu-Olivares (2006) and had not previously been empirically evaluated except for a small simulation study in the original paper. The second test is a new bootstrap procedure that is proposed in the present article. In our simulation study the bootstrap test was the only procedure that adequately maintained Type I error control. Given that ordinal SEM is more sensitive to underlying non-normality than previously assumed, we therefore recommend that researchers run the bootstrap test prior to estimating their models. The test is available in the open-source software package R, and code is provided in the online supplementary material.

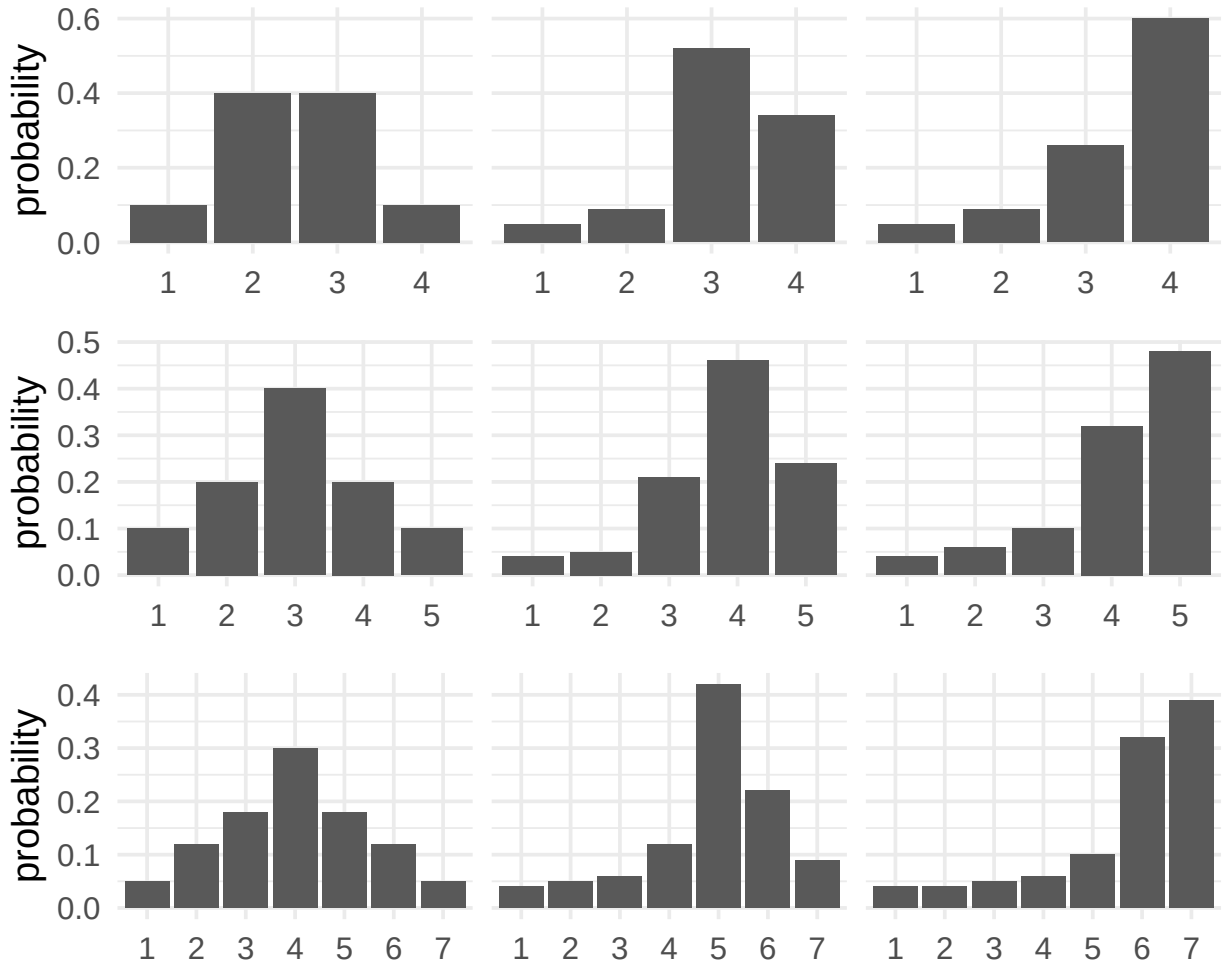


Figure 5. Distribution of the observed variables. The panel rows correspond to $K = 4, 5$ and $K = 7$ categories. The panel columns correspond to conditions of symmetrical, moderately skewed and severely skewed distributions.

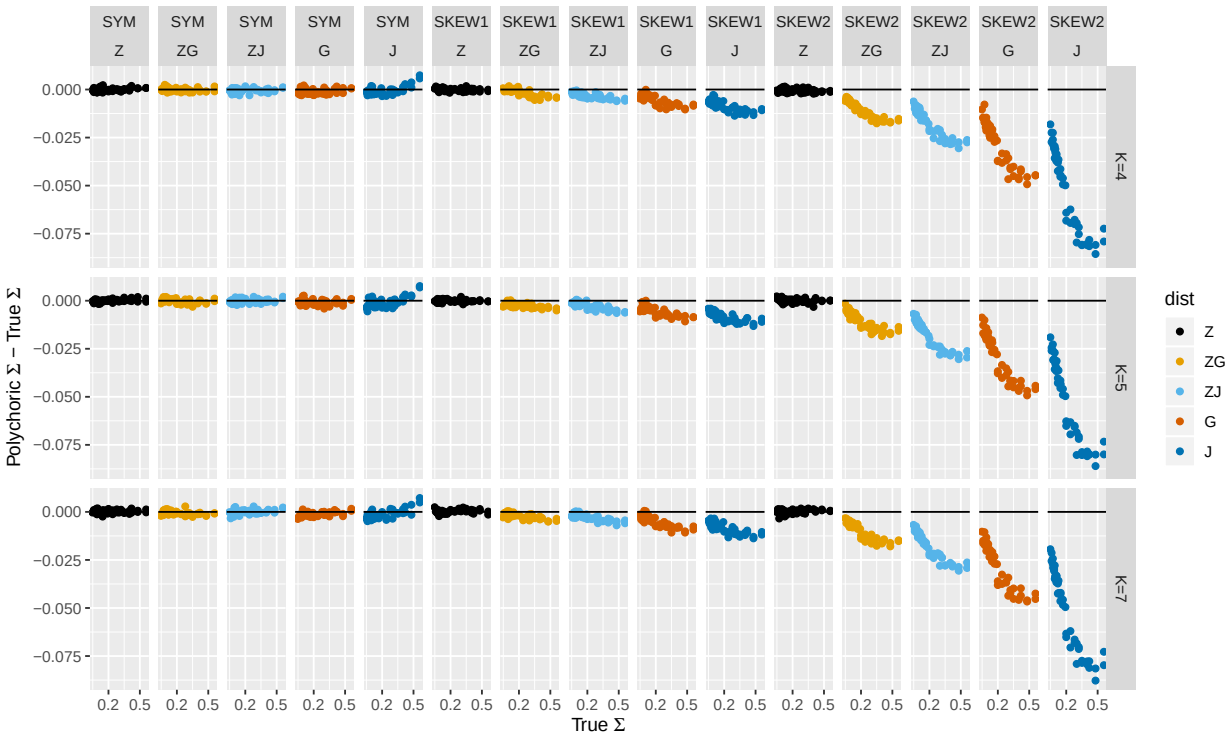


Figure 6. Difference between polychoric and true correlations against true correlations for dimension $p = 10$. The panel rows correspond to $K = 4, 5$ and $K = 7$ categories. The panel columns correspond to conditions of symmetrical, moderately skewed and severely skewed distributions, crossed with five underlying distributions. Z=normal, ZG=normal and Gumbel VITA combined, ZJ=normal and Joe VITA combined, G=Gumbel VITA, J=Joe VITA.

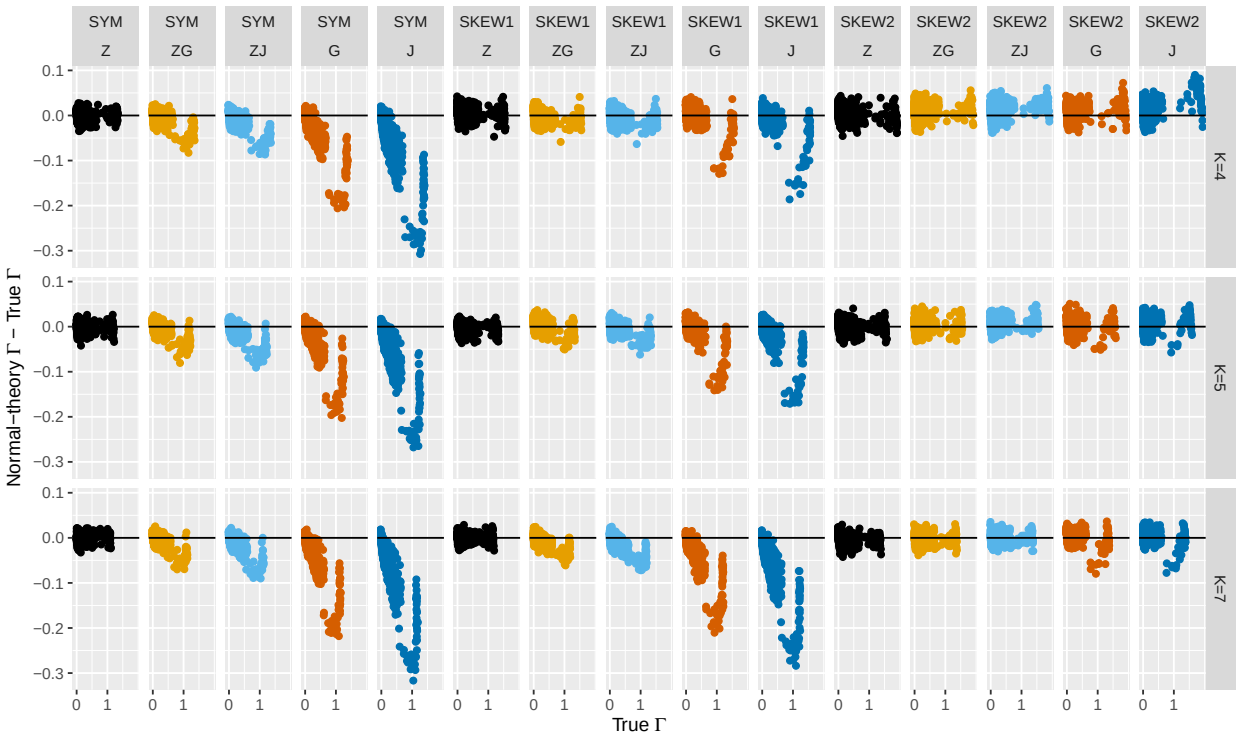


Figure 7. Bias between normal-theory and true values in the covariance matrix of polychoric correlations Γ for dimension $p = 10$. The panel rows correspond to $K = 4$, $K = 5$ and $K = 7$ categories. The panel columns correspond to conditions of symmetrical, moderately skewed and severely skewed distributions, crossed with five underlying distributions. Z=normal, ZG=normal and Gumbel VITA combined, ZJ=normal and Joe VITA combined, G=Gumbel VITA, J=Joe VITA.

n	Cat.	Thr.	Boot	T_S	T_{SS}
200	4	1	0.045	0.119	0.042
		2	0.064	0.149	0.049
		3	0.055	0.122	0.042
	5	1	0.05	0.107	0.028
		2	0.053	0.144	0.041
		3	0.049	0.139	0.037
	7	1	0.073	0.216	0.036
		2	0.052	0.279	0.041
		3	0.057	0.203	0.036
1000	4	1	0.051	0.112	0.049
		2	0.052	0.134	0.055
		3	0.054	0.122	0.055
	5	1	0.06	0.114	0.06
		2	0.06	0.155	0.065
		3	0.053	0.157	0.056
	7	1	0.043	0.12	0.037
		2	0.06	0.186	0.062
		3	0.051	0.188	0.054

Table 4

Type I error rates, dimension $p = 5$. n= sample size. Boot= bootstrap test. T_S and T_{SS}: mean-scaled and mean-and-variance adjusted tests, respectively. Cat.= number of observed categories. Thr.= threshold type, where 1=symmetrical, 2=moderately skewed and 3=strongly skewed. Rejection rates between 2.5% and 7.5% are shaded gray.

p	n	Cat.	Thr.	Boot	T_S	T_{SS}
$p = 10$	200	4	1	0.067	0.13	0.022
			3	0.043	0.13	0.017
		7	1	0.086	0.266	0.002
			3	0.062	0.381	0.009
	1000	4	1	0.06	0.13	0.052
			3	0.051	0.147	0.038
		7	1	0.048	0.118	0.014
			3	0.046	0.195	0.024
$p = 15$	200	4	1	0.076	0.178	0.003
			3	0.043	0.169	0.001
		7	1	0.077	0.315	0
			3	0.079	0.566	0.004
	1000	4	1	0.052	0.14	0.033
			3	0.06	0.155	0.034
		7	1	0.053	0.123	0.004
			3	0.052	0.215	0.017

Table 5

Type I error rates, dimensions $p = 10$ and $p = 15$. n= sample size. Boot= bootstrap test. T_S and T_{SS}: mean-scaled and mean-and-variance adjusted tests, respectively. Cat.= number of observed categories. Thr.= threshold type, where 1=symmetrical and 3=strongly skewed. Rejection rates between 2.5% and 7.5% are shaded gray.

		Dist.	ZG_5		G_5		ZJ_5		J_5	
n	Cat.	Thr.	Boot	T_{SS}	Boot	T_{SS}	Boot	T_{SS}	Boot	T_{SS}
200	4	1	0.06	0.05	0.30	0.25	0.08	0.07	0.77	0.74
		2	0.09	0.08	0.43	0.39	0.18	0.16	0.92	0.92
		3	0.06	0.04	0.12	0.09	0.08	0.06	0.33	0.29
	5	1	0.06	0.04	0.32	0.25	0.12	0.09	0.90	0.84
		2	0.08	0.07	0.43	0.39	0.16	0.15	0.94	0.94
		3	0.08	0.05	0.18	0.14	0.10	0.08	0.58	0.54
	7	1	0.09	0.05	0.24	0.18	0.11	0.07	0.76	0.68
		2	0.10	0.10	0.45	0.42	0.17	0.15	0.92	0.91
		3	0.07	0.04	0.18	0.14	0.10	0.09	0.69	0.66
1000	4	1	0.13	0.12	0.98	0.98	0.36	0.35	1.00	1.00
		2	0.30	0.31	1.00	1.00	0.74	0.75	1.00	1.00
		3	0.12	0.11	0.48	0.47	0.30	0.29	0.99	0.99
	5	1	0.14	0.13	0.99	0.99	0.50	0.48	1.00	1.00
		2	0.22	0.22	1.00	1.00	0.68	0.69	1.00	1.00
		3	0.14	0.14	0.83	0.84	0.46	0.46	1.00	1.00
	7	1	0.11	0.10	0.94	0.93	0.30	0.28	1.00	1.00
		2	0.17	0.17	1.00	1.00	0.52	0.52	1.00	1.00
		3	0.18	0.18	0.94	0.94	0.54	0.55	1.00	1.00

Table 6

Power to detect non-normality, dimension $p = 5$. Dist.=underlying distribution, where ZG=combination of normal and Gumbel VITA, G=Gumbel VITA, ZJ=combination of normal and Joe VITA, and J=Joe VITA. n= sample size. Boot= bootstrap test. T_{SS}=mean-and-variance scaled test. Cat.= number of observed categories. Thr.= threshold type, where 1=symmetrical, 2=moderately skewed and 3=strongly skewed.

			Dist.	ZG		G		ZJ		J	
p	n	Cat.	Thr.	Boot	T_{SS}	Boot	T_{SS}	Boot	T_{SS}	Boot	T_{SS}
$p = 10$	200	4	1	0.07	0.03	0.35	0.16	0.11	0.04	0.88	0.72
			3	0.04	0.01	0.09	0.02	0.06	0.02	0.29	0.16
		7	1	0.09	0.00	0.35	0.03	0.12	0.00	0.86	0.38
			3	0.05	0.01	0.17	0.06	0.06	0.01	0.75	0.57
	1000	4	1	0.14	0.11	0.99	0.99	0.46	0.41	1.00	1.00
			3	0.12	0.10	0.62	0.57	0.28	0.24	1.00	1.00
		7	1	0.12	0.05	0.98	0.95	0.39	0.21	1.00	1.00
			3	0.18	0.14	0.98	0.97	0.62	0.54	1.00	1.00
$p = 15$	200	4	1	0.07	0.00	0.41	0.07	0.13	0.00	0.93	0.64
			3	0.03	0.00	0.08	0.00	0.04	0.00	0.30	0.09
		7	1	0.09	0.00	0.41	0.00	0.16	0.00	0.94	0.09
			3	0.06	0.00	0.15	0.01	0.07	0.01	0.73	0.37
	1000	4	1	0.17	0.10	1.00	1.00	0.52	0.39	1.00	1.00
			3	0.09	0.05	0.68	0.57	0.31	0.22	1.00	1.00
		7	1	0.15	0.02	0.99	0.94	0.45	0.11	1.00	1.00
			3	0.23	0.10	1.00	0.99	0.69	0.48	1.00	1.00

Table 7

Power to detect non-normality, dimensions $p = 10$ *and* $p = 15$*. Dist.=underlying distribution, where ZG=combination of normal and Gumbel VITA, G=Gumbel VITA, ZJ=combination of normal and Joe VITA, and J=Joe VITA. n= sample size. Boot= bootstrap test.* T_{SS}*=mean-and-variance scaled test. Cat.= number of observed categories. Thr.= threshold type, where 1=symmetrical and 3=strongly skewed.*

References

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Unpublished manuscript*. Retrieved from `www.statmodel.com/download/WLSMV_new_chi21.pdf`

Bedford, T., Cooke, R. M., et al. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, *30*(4), 1031–1068.

Bentler, P. (2006). *Eqs 6 structural equations program manual.* Encino, CA: Multivariate Software.

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

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Tech. Rep.). Cambridge University Press.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466–491.

Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*(4), 1066–1083.

Foldnes, N., & Grønneberg, S. (2017a). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–14.

Foldnes, N., & Grønneberg, S. (2017b). The asymptotic covariance matrix and its use in simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–16.

Foldnes, N., & Grønneberg, S. (2019). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika*. (forthcoming)

Foldnes, N., & Olsson, U. H. (2015). Correcting too much or too little? The performance of three chi-square corrections. *Multivariate behavioral research*, *50*(5), 533–543.

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*, 1–17.

Grønneberg, S., & Foldnes, N. (2019). A problem with discretizing Vale-Maurelli in simulation studies. *Psychometrika*, *84* , 554-561.

Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2013). copula: Multivariate dependence with copulas [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=copula` (R package version 0.999-7.)

Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior research methods*, 1–19.

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

Joe, H. (1997). *Multivariate models and multivariate dependence concepts* (Vol. 73). Chapman & Hall/CRC.

Jöreskog, K. G. (2005). *Structural equation modeling with ordinal variables using lisrel.* Technical report, Scientific Software International, Inc., Lincolnwood, IL.

Jöreskog, K. G., & Sörbom, D. (2015). Lisrel 9.20 for windows [computer software]. *Skokie, IL: Scientific Software International.*

Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

Li, C.-H. (2016b). The performance of ml, dwls, and uls estimation with robust corrections in structural equation models with ordinal variables. *Psychological methods*, *21*(3), 369.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57–77.

Maydeu-Olivares, A., Garcia-Forero, C., Gallardo-Pujol, D., & Renom, J. (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology*, *5*(4), 131-136. Retrieved from `https://doi.org/10.1027/1614-2241.5.4.131` doi: 10.1027/1614-2241.5.4.131

Monroe, S. (2018). Contributions to estimation of polychoric correlations. *Multivariate behavioral research*, *53*(2), 247–266.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B., & Muthén, L. (2012). Mplus version 7: User's guide. *Los Angeles, CA: Muthén & Muthén*.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Pearson, K. (1901). Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively. *Philos. Trans. R. Soc. SA*, *196*, 1–47.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2015). On examining the underlying normal variable assumption in latent variable models with categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 581–587.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354.

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Robitzsch, A. (2019). sirt: Supplementary item response theory models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=sirt (R package version 3.4-64)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*(1), 131–151.

Satorra, A., & Bentler, P. (1988). Scaling corrections for statistics in covariance structure analysis (UCLA statistics series 2). *Los Angeles: University of California at Los Angeles, Department of Psychology*.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., & Erhardt, T. (2018). Vinecopula: Statistical inference of vine copulas [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=VineCopula (R package version 2.1.8)

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

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471.