Foldnes, N., & Olsson, U. H. (2019). The Choice of Normal-Theory Weight Matrix When Computing Robust Standard Errors in Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-15. doi:10.1080/10705511.2019.1600408

The choice of normal-theory weight matrix when computing robust standard errors in confirmatory factor analysis

Njål Foldnes and Ulf Henning Olsson

BI Norwegian Business School

Author Note

Correspondence concerning this article should be addressed to N. Foldnes, BI Norwegian Business School. E-mail: njal.foldnes@bi.no

Abstract

Robust standard errors are of central importance in confirmatory factor models. In calculating these statistics a central ingredient is the inverse of the asymptotic covariance matrix of second-order moments calculated under the assumption of normality. Currently, two ways of estimating this matrix are employed in software packages. One approach uses the sample covariance matrix, the other the model-implied covariance matrix. Previous research based on a small confirmatory factor model demonstrated that the latter approach yielded a slight improvement in standard error performance. The present study argues theoretically that the discrepancy between the two approaches increases in models where there are few model parameters relative to $p(p+1)/2$, where p is the number of observed variables. We present simulation results that support this claim, in both small and large correctly specified models, across a large variety of non-normal conditions. We recommend the model-implied covariance matrix for robust standard error computation.

The choice of normal-theory weight matrix when computing robust standard errors in confirmatory factor analysis

Confirmatory factor analysis (CFA) is concerned with the modeling of a vector of observed variables in terms of a system of linear equations relating these observed variables to unobserved variables. As long as these unobserved variables comprise a vector of jointly multivariate normal variables, model parameters may be efficiently estimated with normal-theory maximum likelihood (ML) estimation, and the model tested with the associated likelihood ratio statistic T_{ML}.

However, the normal-theory assumption is seldom met in practice. To take non-normality into account, the current approach in software packages is to provide robust standard errors based on a so-called sandwich matrix (Browne, 1984; Huber, 1967). It is also common to report test statistics that are more robust to non-normality than T_{ML}, with the mean scaling of Satorra and Bentler (1994) as the most well-known example. A central ingredient in the robust formula for standard errors, and in a variety of robust test statistics, is an estimate of a matrix here denoted by W, which coincides with the inverse of the covariance matrix of second-order moments in the case of multivariate normality. For further discussion of alternative ways to obtain robust standard errors and test statistics see Falk (2018); Maydeu-Olivares (2017).

As pointed out by Xia, Yung, and Zhang (2016), there are currently two main ways of estimating W. Both approaches are theoretically correct, and they coincide asymptotically, provided the model is correctly specified. However, with the exception of Xia et al. (2016), there are no empirical studies which investigate whether the two approaches lead to different performance in terms of standard error precision and inference of goodness-of-fit based on robust test statistics. The Monte Carlo studies reported by Xia et al. (2016) were based on a single three-factor model with nine indicators that was estimated at various degrees of misspecification and non-normality. The authors reported on the performance of robust standard errors calculated from the two ways of estimating W. They also

investigated how model fit inference was affected by the choice of $\hat{W}$ when computing the mean-scaled statistic (Satorra & Bentler, 1994), and three fit indices based on this statistic. The results in Xia et al. (2016) revealed a small but consistent performance gap in favor of one version of $\hat{W}$ over the other in terms of standard error precision and model fit. For the fit indices the authors reported only minor differences between the two versions of $\hat{W}$.

Given the widespread use of robust standard errors and test statistics in confirmatory factor analysis, our goal is to expand the study by Xia et al. (2016) in several directions. First we will argue that the modest discrepancy between the two weight matrix approaches reported by Xia et al. (2016) may not be representative for models of a different type than the one employed in Xia et al. (2016). We will explain that, holding other model and the distributional conditions equal, the performance differential in standard error precision and test statistics between the two versions of $\hat{W}$ is likely to become more apparent when the number of free model parameters decreases. It follows that the performance gap may become substantial in large models where there are relatively few parameters to estimate relative to the number of non-redundant elements in the covariance matrix. This explains why Xia et al. (2016) found only minor differences between the two versions of $\hat{W}$, since they used a relatively small model in which relatively many parameters were estimated. We conduct Monte Carlo studies to investigate the degree to which our theoretical analysis is confirmed in various finite-sample conditions. A second extension is the inclusion of a larger variety of non-normal distributions than the single family of distributions used by Xia et al. (2016), in order to study whether the discrepancies between the two weight matrices may be sensitive to certain types of non-normality. A third extension is the inclusion of two more robust test statistics, in addition to the single statistic studied by Xia et al. (2016). In short, our goal is to investigate whether the choice of $\hat{W}$ may affect robust CFA inference to a larger degree than demonstrated in Xia et al. (2016), and to arrive at recommendations for which version of $\hat{W}$ to use.

Robust standard errors and test statistics

Let X be a random p-dimensional vector, with finite fourth-order moments, and with population covariance matrix Σ. Furthermore, let S_n be an unbiased estimator of Σ, based on a random sample of dimension n. The non-duplicated elements of S_n are gathered into the vector $s_n = vech(S_n)$, containing $p^* = \frac{1}{2}p(p+1)$ elements. Similarly we define $\sigma = vech(\Sigma)$. The asymptotic covariance matrix of $\sqrt{n}s_n$ is denoted by Γ. The proposed model contains q parameters and implies a covariance structure $\sigma(\theta)$, where θ is the parameter vector that we assume is differentiable with a Jacobian $p^* \times q$ matrix $\Delta = \frac{\partial \sigma(\theta)}{\partial \theta}$. We assume the model is correct, so that $\sigma(\theta_0) = \sigma$ for some θ_0, and that an estimate $\hat{\theta}_n$ of θ_0 is obtained using an estimator based on multivariate normality, like the ML estimator. Then the asymptotic covariance matrix of $\sqrt{n}\hat{\theta}$ is given by

$$\Omega = (\Delta' W \Delta)^{-1}(\Delta' W \Gamma W \Delta)(\Delta' W \Delta)^{-1}. \tag{1}$$

Here $W = 1/2 D_p'(\Sigma^{-1} \otimes \Sigma^{-1})D_p$, where D_p is the duplication matrix (Magnus & Neudecker, 1999). In the present study we investigate two ways of estimating W:

$$\hat{W}_S = 1/2 D_p'(S_n^{-1} \otimes S_n^{-1})D_p$$

and

$$\hat{W}_\Sigma = 1/2 D_p' \left(\Sigma(\hat{\theta}_n)^{-1} \otimes \Sigma(\hat{\theta}_n)^{-1} \right) D_p.$$

Robust standard error estimates are obtained from eq. (1) when replacing Δ by its estimate $\hat{\Delta}$, and W by either $\hat{W}_S$ or $\hat{W}_\Sigma$. We refer to such standard errors as SE(S) and SE(Σ), respectively.

We next discuss three approximations to the distribution of T_{ML} under violation of multivariate normality. The normal-theory ML estimate $\hat{\theta}_n$ minimizes

$$F_{ML}(\theta) = \log|\Sigma(\theta)| + tr\left(S\Sigma^{-1}(\theta)\right) - \log|S| - p.$$

Under the assumptions of normality and a correctly specified model $T_{ML} = (n-1) \cdot F_{ML}(\hat{\theta})$ is asymptotically distributed as a chi-square with $d := p^* - q$ degrees of freedom. However,

for non-normal data, T_{ML} is asymptotically distributed as a mixture of chi-squares:

$$T_{ML} \quad \xrightarrow{d} \quad \sum_{j=1}^{d} \alpha_j \chi_1^2, \tag{2}$$

where the χ_1^2 are mutually independent chi-squares with one degree of freedom and $\xrightarrow{d}$ stands for convergence in distribution. The α_j are the non-zero eigenvalues of $U\Gamma$ where

$$U = W - W\Delta \left\{\Delta' W \Delta\right\}^{-1} \Delta' W. \tag{3}$$

Based on eq. 2, Satorra and Bentler (1994) proposed to scale T_{ML} by a factor b, yielding

$$T_1 := b^{-1} \cdot T_{ML},$$

where $b := trace(U\Gamma)/d$. The asymptotic mean of T_1 coincides with the mean of the nominal chi-square distribution, namely d. This statistic is regularly reported when there is concern with non-normality of the sample, and is therefore of central interest in CFA. The second statistic, less often reported in the literature, and here denoted by T_2, is closely related to a mean-and-variance-corrected introduced by Satorra and Bentler (1994). T_2 was proposed by Asparouhov and Muthen (2010), and seems to have slightly better performance than the original mean-and-variance-corrected statistic (Foldnes & Olsson, 2015; Savalei & Rhemtulla, 2013). T_2 involves both a scaling and shifting of T_{ML}:

$$T_2 = c_1 \cdot T_{ML} + c_2, \tag{4}$$

where

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

Asymptotically, T_2 has the same mean and variance as the nominal chi-square distribution with d degrees of freedom, so T_2 is theoretically superior to T_1. However, in finite sample conditions both T_1 and T_2 are sensitive to increasing levels of kurtosis in the data, with T_1 tending to produce higher rejection rates than T_2 as kurtosis increases (Foldnes & Olsson, 2015). Note that the parameters b, c_1 and c_2 must be estimated, where $\hat{U}$ is obtained by

inserting $\hat{\Delta}$, and either $\hat{W}_S$ or $\hat{W}_\Sigma$, into eq. (3). When $\hat{W}_S$ is used, we refer to the test statistics as $T_1(S)$ and $T_2(S)$. Similarly, with $\hat{W}_\Sigma$ used, the corresponding test statistics are denoted by $T_1(\Sigma)$ and $T_2(\Sigma)$.

The third test statistic is a member of the newly proposed eigenvalue block averaging (EBA) class of robust test statistics (Foldnes & Grønneberg, 2018). EBA test statistics may be seen as refinements of T_1 where the eigenvalues $\hat{\alpha}_j$ are ordered into two or more blocks and replaced block-wise by their mean values. In this framework, T_1 is the one-block EBA test statistic. In the present study we include the two-block EBA test, here denoted by T_3. We arrange the eigenvalues $\hat{\alpha}_j$ of $\hat{U}\hat{\Gamma}$ in increasing order, and divide them into two equal parts. For the smallest half of the eigenvalues we calculate their average value, $\overline{\alpha}_1$. Similarly, the average of the largest half of the eigenvalues we denote by $\overline{\alpha}_2$. Then, the p-value associated with T_3 is calculated as

$$p = P\left(\sum_{j=1}^{d} \tilde{\alpha}_j Z_j^2 > T_{ML}\right),$$

where $\tilde{\alpha}_k = \overline{\alpha}_1$ for $k = 1, \ldots, \lceil d/2 \rceil$, and $\tilde{\alpha}_l = \overline{\alpha}_2$ for $l = \lceil d/2 \rceil + 1, \ldots, d$, and where the Z_j are independent standard normal variables. Here, $\lceil d/2 \rceil$ is the integer value when $d/2$ is rounded up. Similar to T_1 and T_2, we may calculate T_3 using either $\hat{W}_S$ or $\hat{W}_\Sigma$ in the expression for U, in order to obtain $\hat{U}$ and consequently the estimated eigenvalues. The resulting test statistics are denoted by $T_3(S)$ and $T_3(\Sigma)$, respectively.

To sum up, in order to estimate standard errors and obtain p-values associated with the hypothesis of a correctly specified model from T_1, T_2 and T_3, W and the other matrices that appear in Ω and U must be estimated from sample data and the model specification. Different software packages by default estimate W in different ways before calculating the expressions in equations (1) and (3). For instance, standard errors obtained in M*plus* (Muthén & Muthén, 2010) by ESTIMATOR = MLM or MLMV will be based on $\hat{W}_S$ in eq. (1), while EQS (Bentler, 2006), LISREL (Jöreskog & Sörbom, 2015), SAS/STAT 14.1 ("SAS/STAT 14.1 User's guide", 2015) and lavaan (Rosseel, 2012) use $\hat{W}_\Sigma$ in eq. (1). Similarly, the classical Satorra-Bentler test T_1 and the scaled-and-shifted test T_2 will be

calculated by M*plus* (ESTIMATOR=MLM and MLMV, respectively) using $\hat{W}_S$ in eq. (3), while other software packages like lavaan by default use $\hat{W}_\Sigma$ to obtain an estimate of U.

Distribution of the sample and the model-implied covariance matrices

In this section we discuss how the sampling distributions of the sample covariances s_n and the model-implied covariances $\hat{\sigma}_n$ differ. Distributional differences are likely to be reflected in the distributions of $\hat{W}_S$ and $\hat{W}_\Sigma$, and ultimately in the sampling distributions of the estimates $\hat{\Omega}$ and $\hat{U}\hat{\Gamma}$ that are used to obtain standard errors and robust test statistics.

Although $\hat{\sigma}_n$ and s_n are both vectors of dimension p^*, we will show that the sampling distribution of s_n is more scattered or dispersed than the sampling distribution of $\hat{\sigma}_n$. We first base our discussion on large sample theory, where it is well known that in large samples $\sqrt{n}s_n$ has a distribution which is approximated by a non-degenerate multivariate normal distribution, whose covariance matrix we denote by Γ. In contrast, the large-sample distribution of $\sqrt{n}\hat{\sigma}_n$ is a degenerate normal distribution with covariance matrix $\Delta(\Delta'\Gamma^{-1}\Delta)^{-1}\Delta'$, provided the employed estimator is well specified for the data at hand (e.g., the normal-theory based ML estimator under underlying normality). One way of quantifying scatter in a random vector is to calculate the determinant of its covariance matrix (Wilks, 1932). If we take the determinant of the asymptotic covariance matrix of $\sqrt{n}\hat{\sigma}_n$ the result is zero, since this matrix has rank $q < p^*$. This implies that the scatter or dispersion in $\hat{\sigma}_n$ is reduced, compared to the scatter in s_n.

Informally, the p^*-vector s_n vary in all directions in the p^*-dimensional space in which it lives. In contrast, $\hat{\sigma}_n = \sigma(\hat{\theta}_n)$ is a function of $q < p^*$ estimated parameters, and it spans only a q-dimensional subspace of the p^*-dimensional space in which it resides. It is also noteworthy that the residual vector $\sqrt{n}(s - \hat{\sigma})$ is constrained to a subspace: In large samples this residual vector resides in a subspace of p^* with $p^* - q$ dimensions (Foldnes, Foss, & Olsson, 2011). Hence, s_n enjoys all the p^* degrees of freedom, while $\hat{\sigma}_n$ enjoys only q degrees of freedom. Note that under correct model specification the population

covariance vector σ is included in the q-dimensional space where $\hat{\sigma}_n$ may take on values. This intuitively suggests that $\hat{\sigma}_n$, provided the model is correct, is a better estimate of σ than s_n, because it has less variation than s_n, being constrained to a subspace of p^* that already contains the target σ.

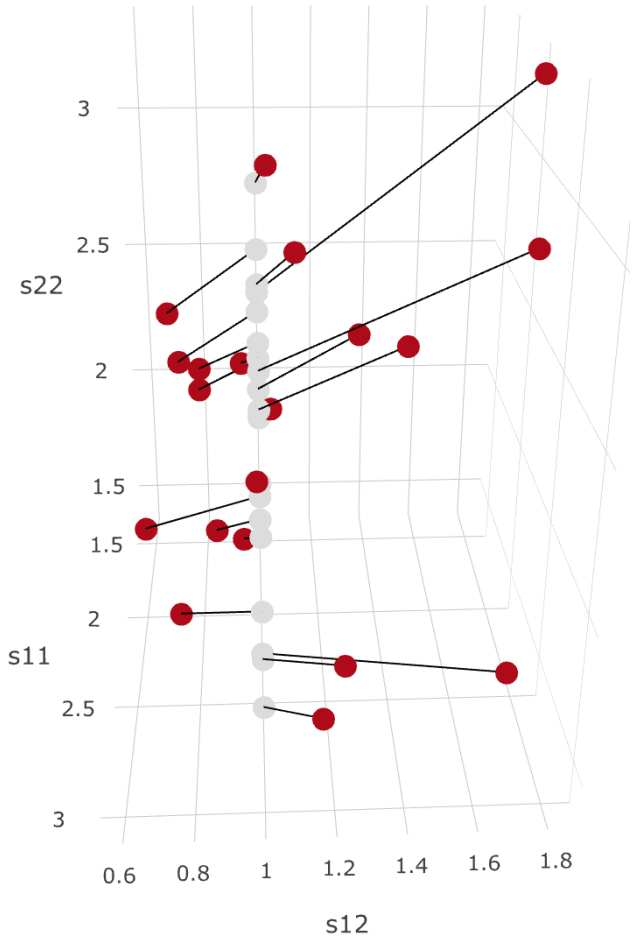

Figure 1. 3d scatterplot of sample covariance vectors (in red) and model-implied covariance vectors (in grey). Pairs s_n and $\hat{\sigma}_n$ estimated from the same sample are joined by a line.

We can visualize this using simulated data from a one-factor analysis model, with $p = 2$ indicator variables X_1 and X_2. In the population model we have unit variances for the factor and the two measurement errors, and the two factor loadings are equal to 1. The resulting population covariance vector is $\sigma = (\sigma_{11}, \sigma_{12}, \sigma_{22}) = (2, 1, 2)$, where σ_{ij} is the population covariance between X_i and X_j. The corresponding sample covariance vector is

$s_n = (s_{11}, s_{12}, s_{22})$. The model we estimate has only $q = 2$ free parameters, namely the unique variances associated with the two variables, while all the other parameters are fixed to their true values. The model-implied covariance vector is then $\hat{\sigma}_n = (1 + \hat{\theta}_1, 1, 1 + \hat{\theta}_2)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the estimated variances of the X_1 and X_2 residuals, respectively. Geometrically, this means that $\hat{\sigma}_n$ is constrained to take on values in a two-dimensional subspace of the three-dimensional space, since $\hat{\sigma}_n$ can only vary along the s_{11} and s_{22} axes, with a s_{12}-coordinate fixed equal to 1. We drew randomly 20 samples of sample size $n = 50$ from a bivariate normal distribution with covariance matrix $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. We estimated the factor model and extracted $\hat{\sigma}_n$ in each sample. The results in the form of a 3D scatterplot are shown in Figure 1. We see that the $\hat{\sigma}_n$ are constrained to lie in the plane defined by $s_{12} = 1$, which is viewed edge-on in the figure, while the s_n are scattered in all three available dimensions.

It follows from the arguments in this section that, keeping the number of observed variables constant, as the number of estimated model parameters decreases, the difference in sampling variability between the model-implied and sample covariance matrices increases. This means that for two nested models, we expect larger differences between SE(S) and SE(Σ) in the most constrained model, which has fewer freely estimated parameters. Another consequence is that the discrepancy between SE(S) and SE(Σ) is expected to increase as model size increases. The reason is that in larger models there are relatively few freely estimated parameters compared to the large number of non-redundant elements in the sample covariance matrix.

In the next section we present a simulation study designed to test these hypotheses. We describe two factor models where one is nested within the other. Our hypothesis then predicts that the former model will exhibit larger discrepancies between SE(S) and SE(Σ) compared to the latter. Also, a model which is much larger than these two models will be presented. Our hypothesis is that for this large model there will be a more pronounced difference between SE(S) and SE(Σ) than observed for the smaller models. In addition to

testing these hypotheses, we are also interested in evaluating the extent to which the choice between SE(S) and SE(Σ) has practical consequence for the quality of parameter inference, by comparing confidence interval coverage rates under the two approaches.

Method

This section provides a detailed description of the proposed models examined in this simulation study, the analyzed sample sizes, and the distributional characteristics evaluated in the simulations in terms of data generation and program implementation. Note that we limit ourselves to the case of correct model specification in the present study.

Models

Xia et al. (2016) simulated exclusively from a 3-factor model, with a total of 9 indicator variables. This model had $q = 21$ free parameters and $d = 24$ degrees of freedom. In the present study we include this model, and denote it by $\mathcal{M}_1^{21}$. In the previous section we argued that the difference between using $\hat{W}_S$ and using $\hat{W}_\Sigma$ in formulas for test statistics and standard errors is more likely to manifest in conditions where q is relatively small compared to p^*. We therefore modified $\mathcal{M}_1^{21}$ by fixing 14 of its free parameters to their population counterparts. More specifically, we fixed six of the nine factor loadings, two of the intrafactor correlations, and six of the nine unique variances. The resulting model has a substantially reduced number of free parameters, with $q = 7$ free parameters to be estimated, compared to $\mathcal{M}_1^{21}$. We refer to this model as $\mathcal{M}_1^{7}$.

Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^{7}$ are relatively small. Confirmatory factor models employed in many simulation studies are typically not much larger, and it is rare to encounter models with more than one hundred degrees of freedom in such studies. Notable exceptions are found in literature examining the effect of large model sizes on test statistics, see, e.g., Shi, Lee, and Terry (2018) and references therein. This body of work is important, given that models with hundreds, or even thousands, degrees of freedom are commonly estimated in many fields of the behavioral sciences. For instance, in personality research commonly used

inventories may contain more than one hundred items (Costa & McCrae, 1992). We therefore included a larger model, referred to as $\mathcal{M}_2$, in the present study. This model has a structure similar to a commonly used measure for the Big Five personality factors, the 60-item NEO Five-Factor Inventory that contains 12 items for each of five factors (Costa & McCrae, 1989). In model $\mathcal{M}_2$ we have $p = 60$ indicators and the $q = 130$ free parameters consists of factor loadings, intrafactor correlations and unique variances. The model degrees of freedom of $\mathcal{M}_2$ is $d = 1700$. In large models, the ratio of the number of free parameters to the degrees of freedom is low, compared to smaller models. Given the discussion in the previous section, we therefore expect the discrepancy between using $\hat{W}_\Sigma$ and using $\hat{W}_S$ to be more pronounced as model size increases.

Data generation

We first discuss distributional conditions. The first distribution condition was that of the multivariate normal distribution. Given the large variety of non-normal distributions, we deem it important to include other types of non-normality than the default type offered in many software packages (Vale & Maurelli, 1983). Also Xia et al. (2016) used the Vale-Maurelli (VM) method for non-normality conditions, but it has its limitations (Astivia & Zumbo, 2018; Foldnes & Grønneberg, 2015). To extend the scope of non-normality, we therefore simulated non-normal data using three recently proposed alternatives to VM, namely the approaches by copula (Mair, Satorra, & Bentler, 2012), by independent generators (Foldnes & Olsson, 2016) and by regular vines (Bedford & Cooke, 2002; Grønneberg & Foldnes, 2017). We refer to these three simulation methods as COP, IG and VITA, respectively.

For models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^{7}$ we used the IG and VITA approaches to generate non-normal data. For the IG distribution, marginal skewness and kurtosis were set to 2 and 7, respectively, values that are often used in simulation studies and considered to represent moderate non-normality. For the VITA distribution, which allows complete

specification of marginals, we set all nine marginals to follow a gamma distribution with shape parameter 6/7 and rate parameter 1. This ensures that the VITA distribution has marginal skewness close to 2 and excess kurtosis equal to 7, similar to the IG distribution. However, VITA differs from IG in having a different copula, which was constructed by employing bivariate Clayton copulas in the regular vine. For models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^7$ data were generated based on setting all factor loadings equal to 0.7, the intrafactor correlations to 0.3, the factor variances equal to 1 and the unique variances equal to 0.51. Two sample sizes, $n = 100$ and $n = 500$, were included. Samples were simulated from three distributions. By fully crossing model, sample size and distributional condition we obtain 12 simulation conditions, in each of which 2000 samples were generated.

The distributional conditions for $\mathcal{M}_2$ were the multivariate normal, one VM distribution and two COP distributions. For the VM distribution we specified skewness 2 and excess kurtosis 7 in each of the 60 marginals. In the first COP distribution, we started with a Gumbel copula with parameter 5, and applied marginal gamma distributions with shape parameter 6/7. Then the simulated data were linearly transformed to obtain a correct covariance matrix. We refer to this distribution as COP1. For the second COP distribution, COP2, the same marginals were employed, but we started with a t-copula with parameter 0.5. For both COP1 and COP2 the linear transformation was calculated using a warm-up sample size of $n = 6 \cdot 10^6$. The correctness of COP1 and COP2 was checked using the asymptotically distribution-free test of correct covariance matrix specification proposed by Mair et al. (2012).

For model $\mathcal{M}_2$, to improve external validity of the simulation results, population parameters were set to values close to those observed in empirical studies using the NEO instrument. The big five factors are weakly correlated, so we fixed intrafactor correlations to 0.3. Also, factor loadings were set to values that ensured that the population coefficient alpha for each of the five scales equaled $0.7, 0.7, 0.8, 0.8$ and 0.9. These alpha coefficients are in the same range as those observed in empirical studies (McCrae & Costa, 2007). Two

sample sizes, $n = 500$ and $n = 2000$, were included. Samples were simulated from four distributions. By fully crossing sample size and distributional condition we obtain 8 simulation conditions, in each of which 2000 samples were generated.

Data analysis

Data generation and model estimation were conducted in the R programming environment, using packages lavaan (Rosseel, 2012), copula (Kojadinovic & Yan, 2010) and VineCopula (Schepsmeier et al., 2018). Simulations for $\mathcal{M}_2$ were conducted on the Abel computer cluster, owned by the University of Oslo and Uninett/Sigma2.

Each simulated data set was estimated using normal-theory ML estimation. In the data analysis all 2000 replications were used for all study design cells, because no problems with estimation convergence occurred. Standard errors and test statistics were then calculated in two ways, using $\hat{W}_\Sigma$ and $\hat{W}_S$. In each simulation condition we calculated the following outcome variables:

1. For each estimated parameter, the mean value of the estimated standard errors SE(S) and SE(Σ). Empirical standard error for each estimated parameter. For each estimated parameter, the confidence interval coverage rate at the 95 % level of confidence, based on both SE(S) and SE(Σ).

2. Rejection rate at the $\alpha = 0.05$ level of significance for the test statistics $T_i(S)$ and $T_i(\Sigma)$, for $i = 1, 2, 3$.

In the next section we report our results. For the presentation of standard error precision and confidence interval coverage previous related research has used tables, see, e.g., Falk (2018); Maydeu-Olivares (2017); Xia et al. (2016). Due to large table size, this only allows the reporting of the results for a few parameters. In the present study, we instead used graphs to depict our results for standard errors. In our opinion, this leads to better interpretability of the simulation outcomes, a great advantage being that standard

errors and coverage rates for *all* free parameters in the model may be included in the reporting. Hence, using graphs facilitates the reporting of more data from the simulation studies, but at the expense of decimal precision within each simulation cell. For test statistic rejection rates we employ tables.

Results

Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^7$

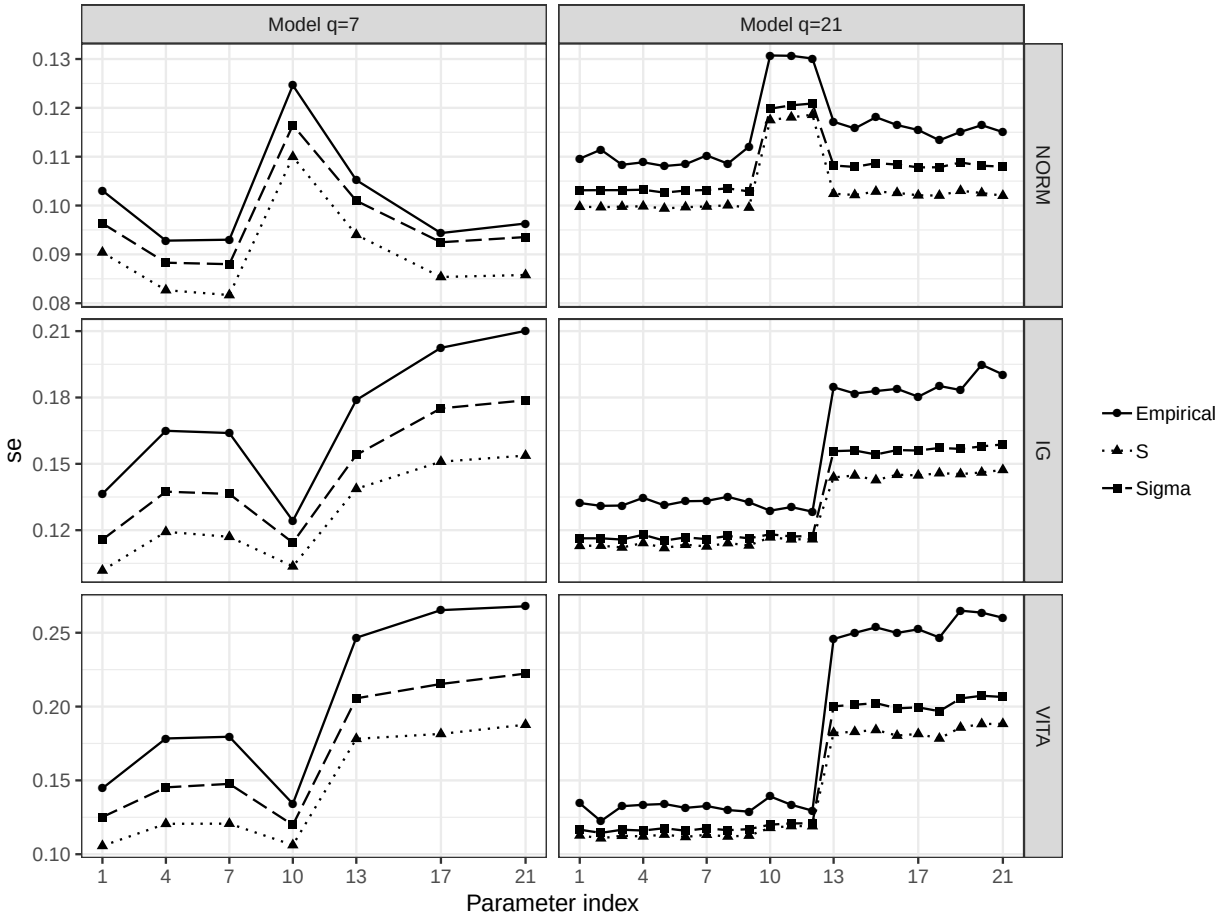

Figure 2. Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^7$ at sample size $n = 100$, in right and left column of panels, respectively. NORM=Normal distribution. IG=Independent generator distribution. VITA= Regular vine distribution. Empirical= The empirical standard error. S= Mean of SE(S). Sigma= Mean of SE(Σ).

For sample size $n = 100$ the means of the estimated standard errors and empirical standard errors are presented in Figure 2. The left column of panels concern $\mathcal{M}_1^7$, where only 7 parameters were estimated. The right column of panels concerns $\mathcal{M}_1^{21}$, with 21 parameter indices i plotted on the x-axis. The first group of parameters are factor loadings ($1 \leq i \leq 9$). The next group ($10 \leq i \leq 12$) refers to factor correlations, while the last group ($13 \leq i \leq 21$) corresponds to the nine unique variances. For all parameters, the empirical standard error is larger than both SE(S) and SE(Σ). However SE(Σ) consistently outperforms SE(S) across all parameters, distributions and models. Notably, the superior performance of SE(Σ) relative to SE(S) is accentuated in the model with only $q = 7$ free parameters. As expected, we observe an increase in empirical standard errors under non-normality. Interestingly, while the factor correlations were estimated with the lowest precision among the parameters under normality, it was the unique variances that were least precisely estimated under both non-normal distributions.

For sample size $n = 500$ the standard error results are given in Figure 3. As expected, the empirical standard errors are lower than for the $n = 100$ case. Similar to the $n = 100$ case, SE(S) and SE(Σ) consistently underestimate the empirical standard error across all parameters, but the bias is reduced compared to the $n = 100$ case. We again observe that SE(Σ) is consistently a better estimator of the empirical standard error than is SE(S). This performance gap is again larger in the model with the smallest q, namely $\mathcal{M}_1^7$. However, discrepancy in performance between SE(S) and SE(Σ) is smaller in the $n = 500$ case compared to the $n = 100$ case.

To better interpret the practical significance of these findings, we present for the $n = 100$ case a plot of confidence interval coverage rates in Figure 4. The coverage rate is the proportion of times the confidence interval based on the calculated standard error contains the true population value. This rate was calculated with a 95 % confidence level, so ideally the coverage rate should be close to 0.95. As expected, given the underestimation of standard errors, the coverage rates are all below the nominal 95 % level. However,

coverage rates are consistently better for SE(Σ) compared to SE(S), across both models, all three distributions and all 21 parameters. The gap in performance is most marked for the model with lowest q. For instance, the coverage rate for the first loading parameter in $\mathcal{M}_1^7$ under VITA is 0.83 when based on SE(S), compared to 0.89 when based on SE(Σ). In Figure 5 are presented the coverage rates for $n = 500$. Here the discrepancy in performance between SE(S) and SE(Σ) is less pronounced compared to the $n = 100$ condition, but still SE(Σ) coverage rates consistently outperform those of SE(S). Although in many conditions the slight gap in coverage rates may not seem of practical importance, we note that for the model with smallest q there are some parameters and distributions where the gap is still notable.

Next we proceed to consider test statistic rejection rates. Table 1 presents rejection rates of the test statistics T_1, T_2 and T_3, based on both $\hat{W}_\Sigma$ and $\hat{W}_S$. Under the normal and IG distributions there is overall little difference between the two versions $T_i(\Sigma)$ and $T_i(S)$ across all three statistics, $i = 1, 2, 3$. However, a gap appears in the smallest model ($q = 7$) under IG, when $n = 100$. For the VITA distribution there are larger discrepancies between $T_i(\Sigma)$ and $T_i(S)$, especially for T_1 and $q = 7$. Overall, the discrepancies are more pronounced at the smallest model size. Compared to the clear-cut situation for standard errors, it is less clear for test statistics whether $\hat{W}_\Sigma$ or $\hat{W}_S$ should be preferred. In the case of the Satorra-Bentler test T_1, the tendency to overreject is mitigated with $\hat{W}_\Sigma$ compared to $\hat{W}_S$, under the non-normal distributions. For T_2 the tendency to underreject seems to be modestly mitigated by using $\hat{W}_S$ instead of $\hat{W}_\Sigma$. For T_3 the tendency to overreject seems to be mitigated by preferring $\hat{W}_\Sigma$.

Model $\mathcal{M}_2$

For sample size $n = 500$ the means of the estimated standard errors and the empirical standard error are presented in Figure 6. There are 130 parameters in the model, whose indices i plotted on the x-axis. The first group of parameters are factor loadings

($1 \leq i \leq 60$). The next group ($61 \leq i \leq 70$) refers to factor correlations among the five factors, while the last group ($71 \leq i \leq 130$) corresponds to the sixty unique variances.

Under multivariate normality, the mean of SE(Σ) closely matches the empirical standard errors across all parameters. SE(S) does not perform as well, consistently underestimating the empirical standard errors. As expected, standard errors are larger under the non-normal distributions. In terms of empirical standard errors, the non-normality embedded in VM, COP1 and COP2 is increasingly challenging for the ML estimator.

Of crucial interest is the discrepancy between SE(S) and SE(Σ). An important observation, visible across all parameters and distributions in Figure 6, is the poor performance of SE(S) relative to SE(Σ). For the VM distribution the performance gap is especially prominent for the unique variances, whose empirical standard errors are close to 0.14, while the SE(S) are close to 0.1. The performance gap between SE(S) relative to SE(Σ) is even more accentuated for COP1 and, especially, for COP2. COP2 represents a more severe non-normality than VM and COP1, where there are large gaps between SE(S) and SE(Σ) for both factor loadings and unique variances. The practical implication of this performance gap is investigated with confidence interval coverage rates plots in Figure 7. Even under multivariate normality there is a practical difference in coverage rates, for the unique variances the SE(Σ) coverage rates are close to the 95 % nominal rate, while the SE(S) coverage rates are typically close to 91 %. This gap widens markedly under the non-normal distributions, where SE(Σ) coverage rates are consistently maintained at 90 % or above. In contrast, SE(S) coverage rates are lower, especially under COP2. For COP2 the SE(S) coverage rates fall well below 70% for most parameters.

At the large $n = 2000$ sample size, standard errors and coverage rates are plotted in Figures 8 and 9 and reflect the findings for $n = 500$. As expected, empirical standard errors are smaller for $n = 2000$ compared to the $n = 500$ condition. It is noteworthy that while SE(Σ) closely approximates the empirical standard errors, again SE(S) exhibits a

downward bias, especially in the non-normal conditions. This is again most notable for COP2, where standard errors of factor loadings and unique variances are markedly underestimated. Interestingly, for VM and COP1, SE(S) for the factor loadings (parameter index $i \leq 60$) are not as bad as SE(S) for the unique variances. We observe, in accordance with the asymptotical nature of eq.(1), that both SE(S) and SE(Σ) lead to more acceptable coverage rates for the $n = 2000$ case compared to the $n = 500$ case. However, even at this large sample size, there is a practical difference in performance between SE(S) and SE(Σ) in all the three non-normal conditions. Under COP2, coverage rates for the unique variances are close to 85% when based on SE(S), clearly inferior to SE(Σ) coverage rates, which are close to 93%.

Rejection rates for test statistics with model $\mathcal{M}_2$ are presented in Table 2. Under multivariate normality the performance differences between $T_i(\Sigma)$ and $T_i(S)$, for $i = 1, 2, 3$, are negligible. Under non-normality rejection rates associated with both versions of T_1 are too high under VM and COP1, although $T_1(\Sigma)$ comes closer to the nominal rejection level than $T_1(S)$. However, this situation is reversed under COP2, where indeed all test statistics based on $\hat{W}_\Sigma$ nearly always accept the model. This is in stark contrast to basing the same statistics on $\hat{W}_S$, which leads to rejection in almost all instances for T_1 and T_3. There are also marked differences between $T_1(\Sigma)$ and $T_1(S)$ for VM and COP1. To summarize, for T_1 rejection happens more often when based on $T_1(S)$ compared to $T_1(\Sigma)$, and the latter version is to be preferred. However, note that both $T_1(\Sigma)$ and $T_1(S)$ deliver inacceptable Type I error controls. For T_2 the situation is the opposite, rejection happens more often with $T_2(\Sigma)$ compared to $T_2(S)$. Also for T_2 the Type I error control is inacceptable, and there is no clear version to be preferred. For T_3 the model is more often rejected with $T_3(S)$ compared to $T_3(\Sigma)$, and the latter version is preferable under VM and COP1, at least for the largest sample size. Under COP2 both the T_3 versions are inacceptable.

Discussion

The Monte Carlo investigation in Xia et al. (2016) revealed that SE(Σ) consistently outperformed SE(S), and also that $T_1(\Sigma)$ demonstrated a slightly better performance than $T_1(S)$. However, the differences reported were minor. The purpose of the present study was to investigate whether this result generalizes to other types of models and underlying non-normal distributions. We also wanted to confirm our theoretical expectation that with fewer parameters to be estimated, the discrepancy between SE(Σ) and SE(S) increases, and that for large models the discrepancy will tend to be more pronounced.

First, our findings were in accordance with Xia et al. (2016) on the preferred way of calculating robust standard errors, and confirmed that SE(Σ) outperforms SE(S). However, we also found that the discrepancy between the two alternatives may be far from minor. In many cases the choice of weight matrix $\hat{W}$ has a substantial effect on the quality of robust CFA inference.

We provided theoretical explanation for the fact that the smaller q is relative to p^*, the more pronounced the gap in sampling distribution between $\hat{W}_\Sigma$ and on $\hat{W}_\Sigma$ will become, and this gap may transfer to robust statistics like standard errors. In order to demonstrate that the ratio q/p^* is an important predictor of the discrepancy of outcomes based on $\hat{W}_S$ and $\hat{W}_\Sigma$, we included in our study two models in addition to the model $\mathcal{M}_1^{21}$ used by Xia et al. (2016). In the smallest model, $\mathcal{M}_1^7$, the q/p^*-ratio was 7/45, compared to 21/45 for in $\mathcal{M}_1^{21}$. In the large model, $\mathcal{M}_2$ the q/p^*-ratio was even smaller than in $\mathcal{M}_1^7$, namely 130/1830. Our conjecture was that in models $\mathcal{M}_1^7$ and $\mathcal{M}_2$, the discrepancies between SE(Σ) over SE(S), would be more pronounced than they were in model $\mathcal{M}_1^{21}$. We also expected larger discrepancies in test statistic performances in these models. This was confirmed in the simulation studies.

Our findings revealed that the difference between using $\hat{W}_S$ and using $\hat{W}_\Sigma$ in robust standard error calculation may become quite dramatic. Across all models, parameters, sample sizes and distributional conditions confidence interval coverage rates based on

SE(Σ) were found to perform better than SE(S). This gap increased with increasing non-normality, with decreasing sample size, and with a decrease in the ratio q/p^*. Considering models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^{7}$, the largest discrepancies in coverage rates between SE(S) and SE(Σ) occurred in the smallest model, $\mathcal{M}_1^{7}$, at the smallest sample size $n = 100$, under the VITA distribution. In this condition, across seven estimated parameters, the mean coverage rates based on SE(S) and SE(Σ) were 79.6% and 85.1%, respectively. For the larger model $\mathcal{M}_1^{21}$ in the same $n = 100$/VITA condition, the respective coverage rates were 85.5% and 83.8%. So increasing the number of free parameters from $q = 7$ to $q = 21$ reduced the discrepancy between SE(Σ) and SE(S), rendering it almost negligible. Our findings of a consistent but small discrepancy between SE(Σ) and SE(S) for model $\mathcal{M}_1^{21}$ is in accordance with the results in Xia et al. (2016). The discrepancy between the two ways of calculating standard errors was found to be even more pronounced in a large five-factor model, with 60 indicators. This model, $\mathcal{M}_2$, is similar in size to models commonly estimated in personality research. We calibrated the population values to resemble values reported in big five research, and found that the superior performance of SE(Σ) over SE(S) again was most notable at the smallest sample size under a severely non-normal distribution. Under the COP2 distribution, with $n = 500$, the mean coverage rate across all 130 parameters was 90.9% for SE(Σ), compared to the unacceptably low mean coverage rate of 65.7% for SE(S).

We also evaluated rejection rates for three robust test statistics, each of which was calculated using either $\hat{W}_S$ or $\hat{W}_\Sigma$. We observed that the resulting discrepancies were larger in $\mathcal{M}_1^{7}$ than in $\mathcal{M}_1^{21}$, especially at the smallest sample size. For the Satorra-Bentler mean-scaled statistic, the tendency to overreject under non-normality was reduced by using $\hat{W}_\Sigma$ instead of $\hat{W}_S$, which echoes findings in Xia et al. (2016). For the scaled-and-shifted statistic and the eigenvalue block averaging statistic, we found that the choice of weight matrix had only a minor impact on Type I error control. It is well known that in large models, the Satorra-Bentler statistic tends to overreject true models, even under normal

data (Herzog, Boomsma, & Reinecke, 2007). We confirmed these findings under model $\mathcal{M}_2$. Although $\hat{W}_\Sigma$ yielded a better performance of the Satorra-Bentler statistic than $\hat{W}_S$, the rejection rates were still unsatisfactory. The scaled-and-shifted statistic almost never rejected the true model when based on $\hat{W}_\Sigma$, and exhibited better performance when based on $\hat{W}_S$. For the eigenvalue block-averaging statistic, performance was best under $\hat{W}_\Sigma$. For the most severe non-normality condition, COP2, none of the three statistics came close to acceptable performance. All three statistics showed a large discrepancy, with virtually zero rejection rate when based on $\hat{W}_\Sigma$, while rejecting almost all models when based on $\hat{W}_S$. In sum, we may recommend $\hat{W}_\Sigma$ for the Satorra-Bentler test, $\hat{W}_S$ for the scaled-and-shifted test, and $\hat{W}_\Sigma$ for the EBA test. However, it should be noted that none of these statistics had adequate performance under $\mathcal{M}_2$, across the range of distributions included in the present study. Better statistics for large models and small n, based on Bartlett or Swain corrections, are discussed in Shi et al. (2018).

Our study has of course limitations. First, we considered only factor analytical models. That is, we did not investigate structural equation modeling extensions such as growth or multilevel models. Secondly, our models were all correctly specified. What happens in the more realistic scenario of an incorrectly specified model? We conjecture that the model-implied covariance matrix should still be used as long as the model is approximately correct. In case of a severely incorrect model the usefulness of standard error estimation may be limited.

Conclusion

We investigated whether the sample or the model-implied covariance matrix should serve as basis for calculating the asymptotic covariance matrix under assumed normality. This matrix is an important ingredient in commonly used formulas for so-called robust CFA inference. By theoretical arguments, and by simulating from three factor analytical models, across several kinds of non-normal distributions, we found that standard error

calculation should be based on the model-implied covariance matrix. This is especially important at small sample sizes with non-normal data, and when there are few estimated parameters in the model relative to the number $p(p+1)/2$ of non-redundant elements in the covariance matrix (e.g., in large models with more than, say, 40 indicator variables).

Three robust test statistics were included, for which we have less clear recommendations. The scaled and the EBA statistic should be based on the model-implied matrix, while for the scaled-and-shifted statistic the sample covariance matrix seemed to offer best performance. However, all three statistics exhibited inadequate Type I error control in distributional conditions furthest removed from multivariate normality.

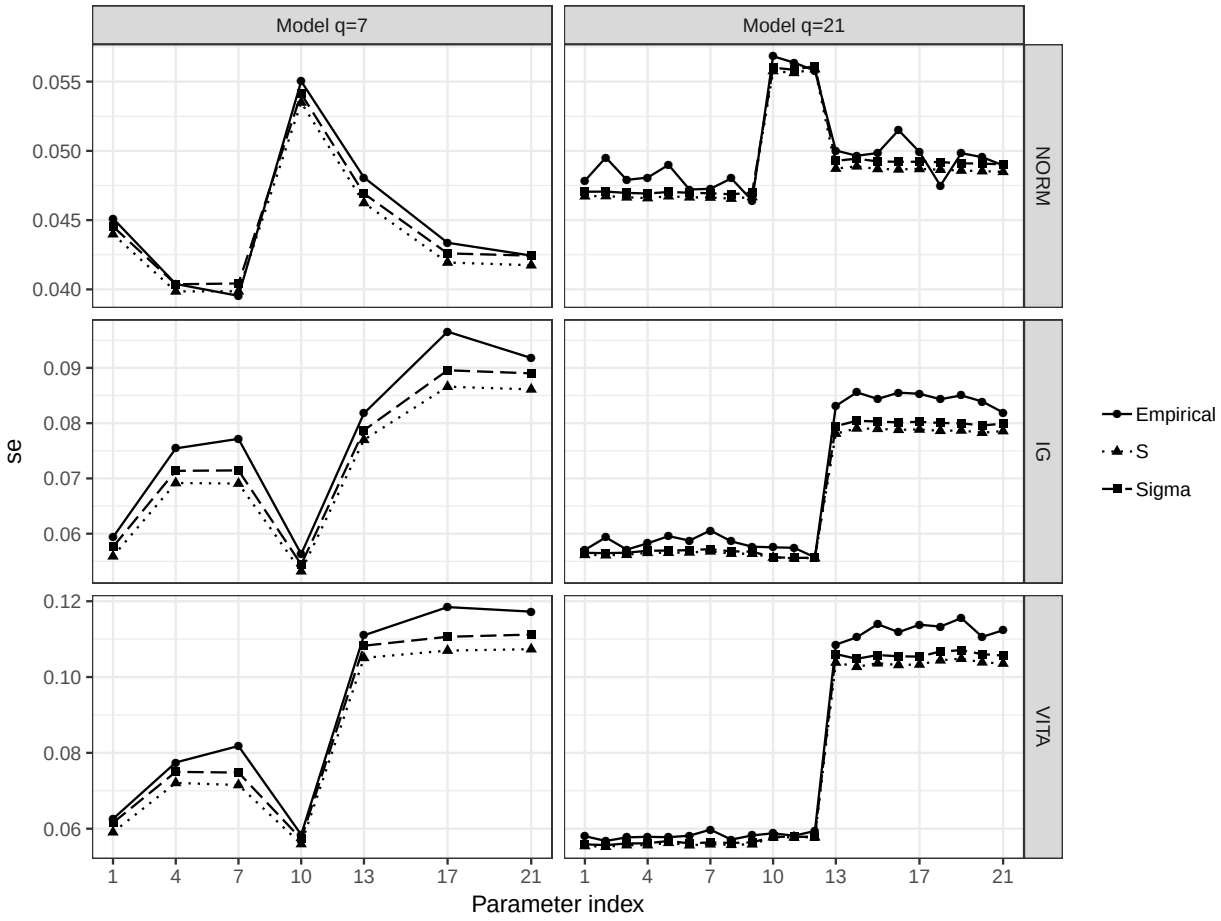

Figure 3. Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^7$ at sample size $n = 500$, in right and left-hand columns of panels, respectively. NORM=Normal distribution. IG=Independent generator distribution. VITA= Regular vine distribution. Empirical= The empirical standard error. S= Mean of SE(S). Sigma= Mean of SE(Σ).

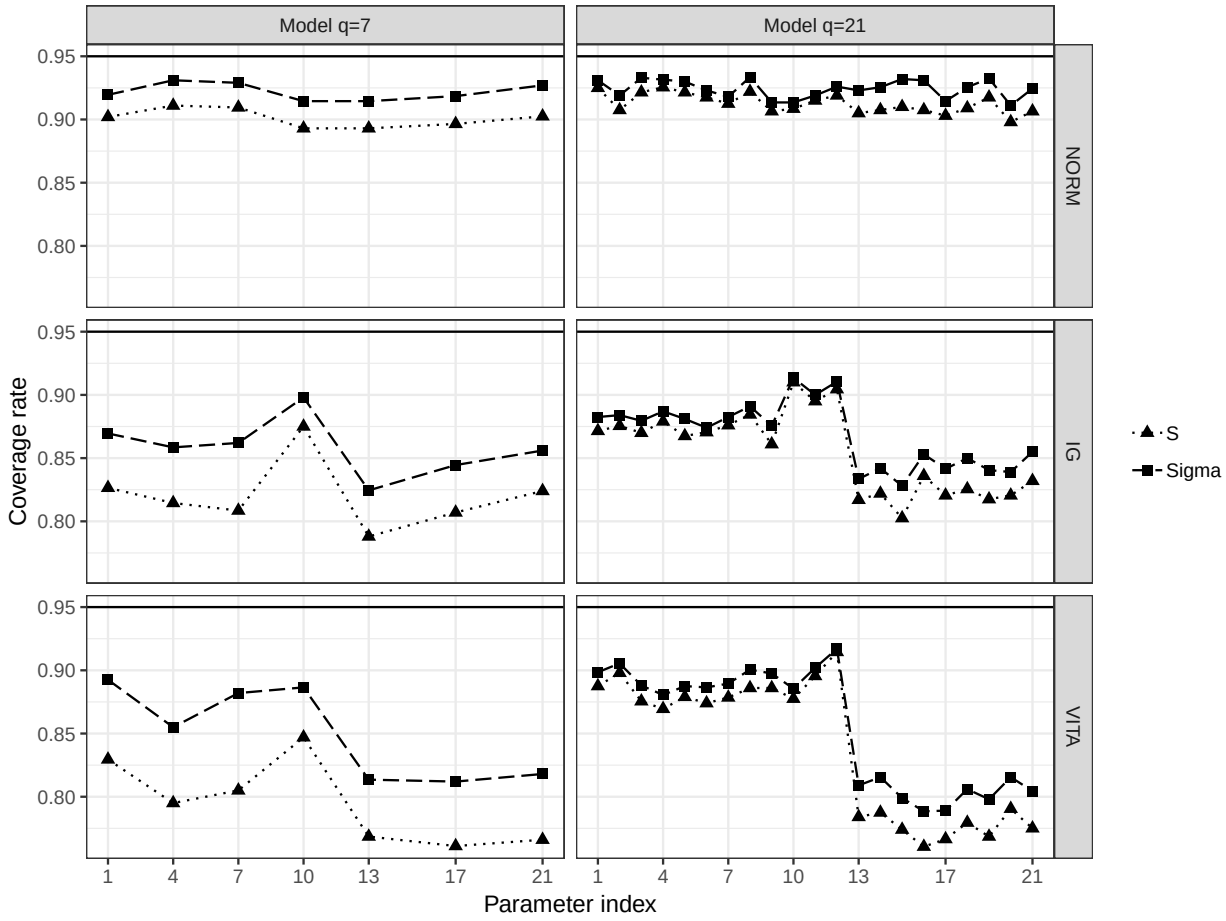

Figure 4. Confidence interval coverage rates for Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^{7}$ at sample size $n = 100$. NORM=Normal distribution. IG=Independent generator distribution. VITA= Regular vine distribution. S= Coverage rate based on SE(S). Sigma= Coverage rate based on SE(Σ).

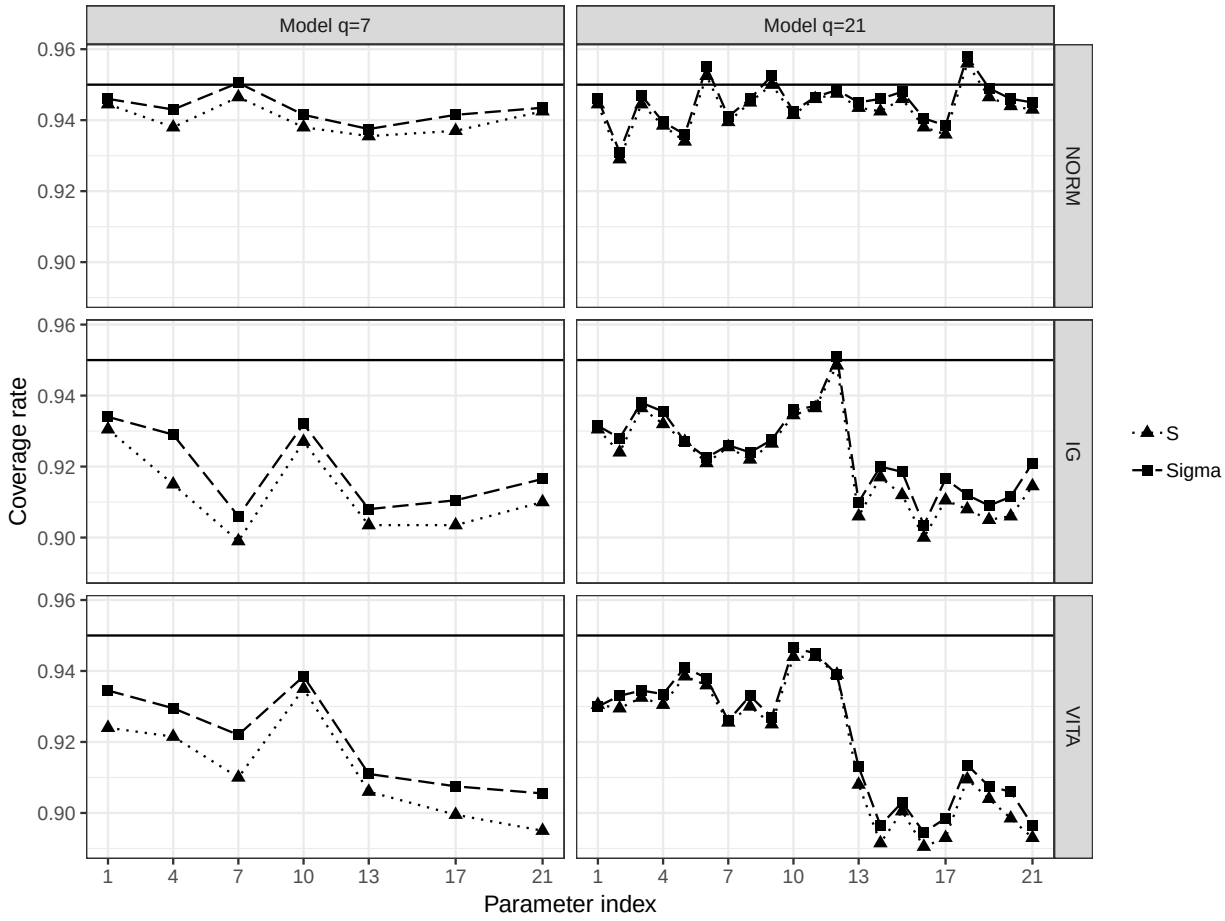

Figure 5. Confidence interval coverage rates for Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^7$ at sample size $n = 500$. NORM=Normal distribution. IG=Independent generator distribution. VITA= Regular vine distribution. S= Coverage rate based on SE(S). Sigma= Coverage rate based on SE(Σ).

Dist	n	q	$T_1(\Sigma)$	$T_1(S)$	$T_2(\Sigma)$	$T_2(S)$	$T_3(\Sigma)$	$T_3(S)$
Normal	100	7	0.14	0.11	0.07	0.07	0.09	0.08
		21	0.08	0.09	0.05	0.06	0.06	0.07
	500	7	0.06	0.06	0.05	0.06	0.05	0.06
		21	0.06	0.06	0.05	0.05	0.05	0.05
IG	100	7	0.22	0.26	0.10	0.11	0.16	0.18
		21	0.09	0.10	0.03	0.05	0.05	0.06
	500	7	0.13	0.14	0.06	0.07	0.09	0.11
		21	0.06	0.06	0.03	0.03	0.04	0.04
VITA	100	7	0.28	0.34	0.13	0.14	0.21	0.24
		21	0.08	0.12	0.01	0.03	0.03	0.05
	500	7	0.13	0.15	0.06	0.06	0.10	0.11
		21	0.06	0.06	0.03	0.03	0.04	0.04

Table 1

Models $\mathcal{M}_1^{21}$ and $\mathcal{M}_1^{7}$ rejection rates calculated at the $\alpha = 0.05$ level of significance. Normal= multivariate normal distribution. IG= Independent generator distribution. VITA= Regular vine distribution. n= sample size. q=Number of free parameters estimated. T_1=Satorra-Bentler test. T_2=Scaled-and-shifted test. T_3=Eigenvalue block averaging with 2 blocks. For each test results using both $\hat{W}_\Sigma$ and $\hat{W}_S$ are presented.

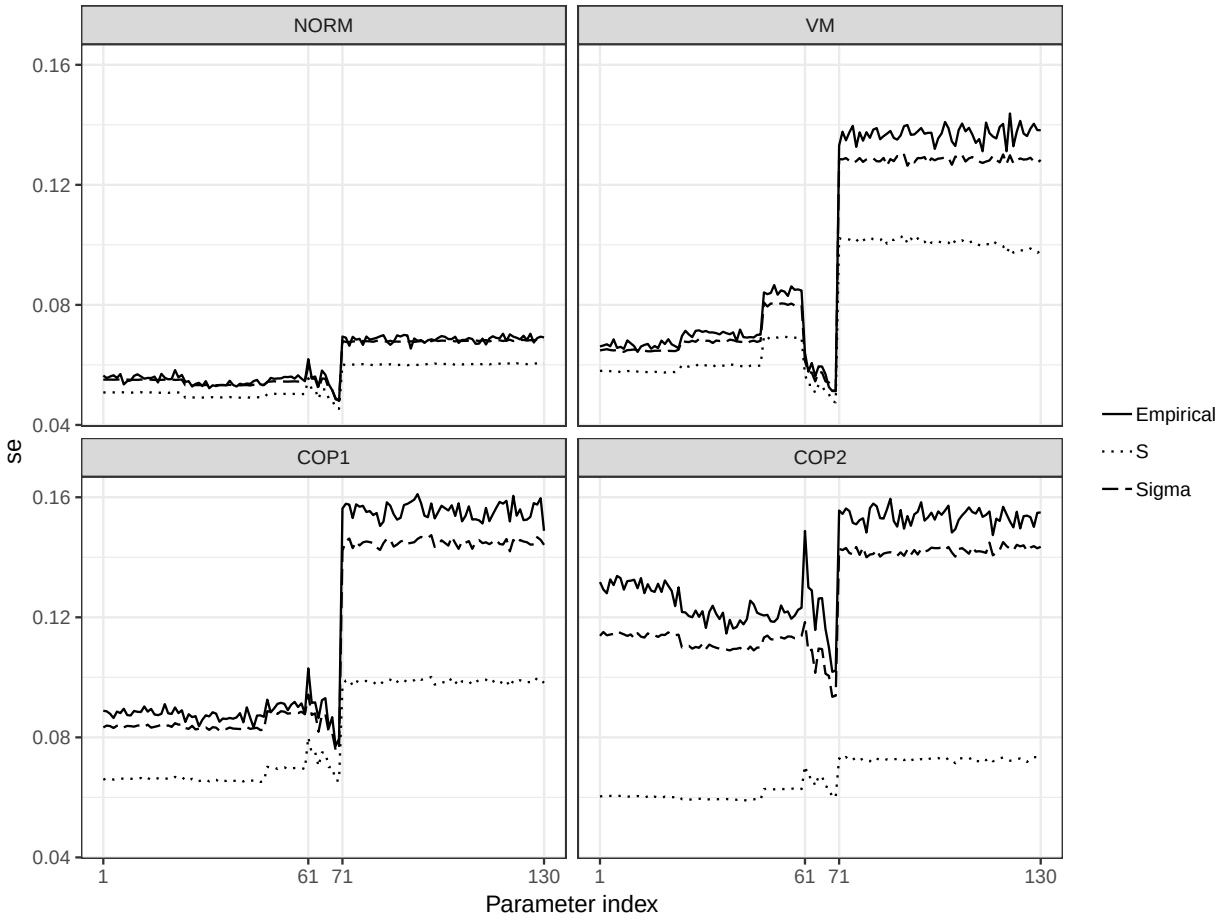

Figure 6. Model $\mathcal{M}_2$ at sample size $n = 500$. NORM=Normal distribution. VM=Vale-Maurelli distribution. COP1= Gumbel copula distribution. COP2= Student's t copula distribution. Empirical= The empirical standard error. S= Mean of SE(S). Sigma= Mean of SE(Σ).

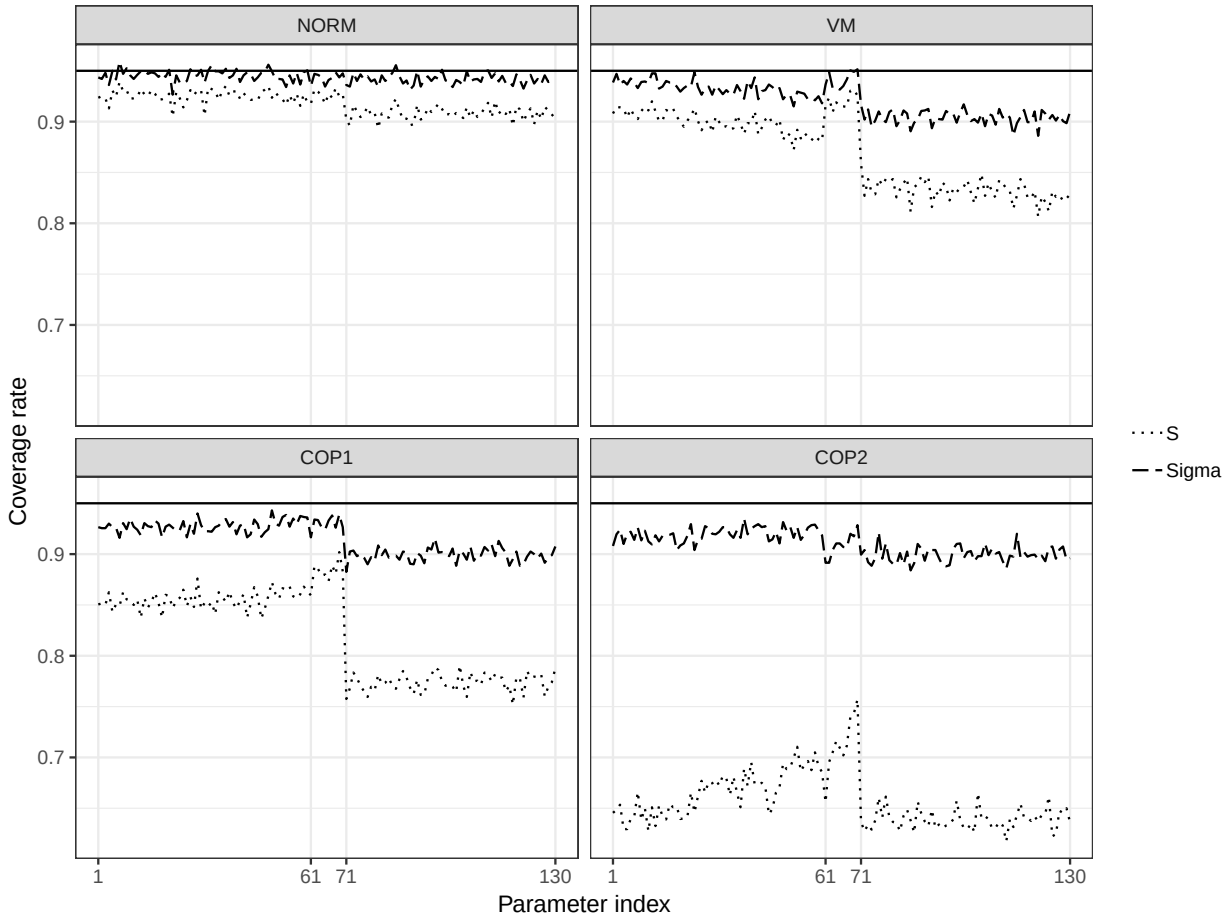

Figure 7. Confidence interval coverage rates for Model $\mathcal{M}_2$ at sample size $n = 500$. NORM=Normal distribution. VM=Vale-Maurelli distribution. COP1= Gumbel copula distribution. COP2= Student's t copula distribution. S= Coverage rate based on SE(S). Sigma= Coverage rate based on SE(Σ).

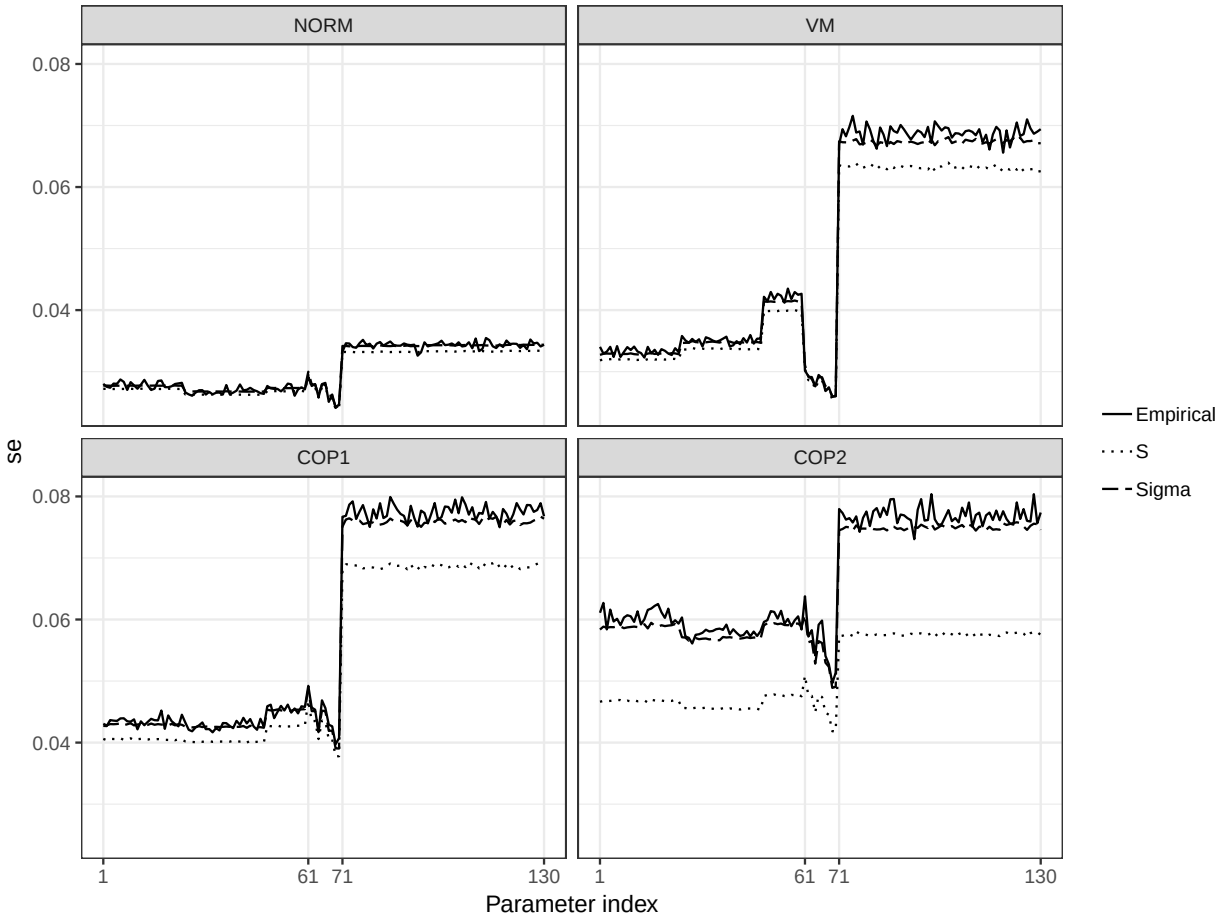

Figure 8. Model $\mathcal{M}_2$ at sample size $n = 2000$. NORM=Normal distribution. VM=Vale-Maurelli distribution. COP1= Gumbel copula distribution. COP2= Student's t copula distribution. Empirical= The empirical standard error. S= Mean of SE(S). Sigma= Mean of SE(Σ).

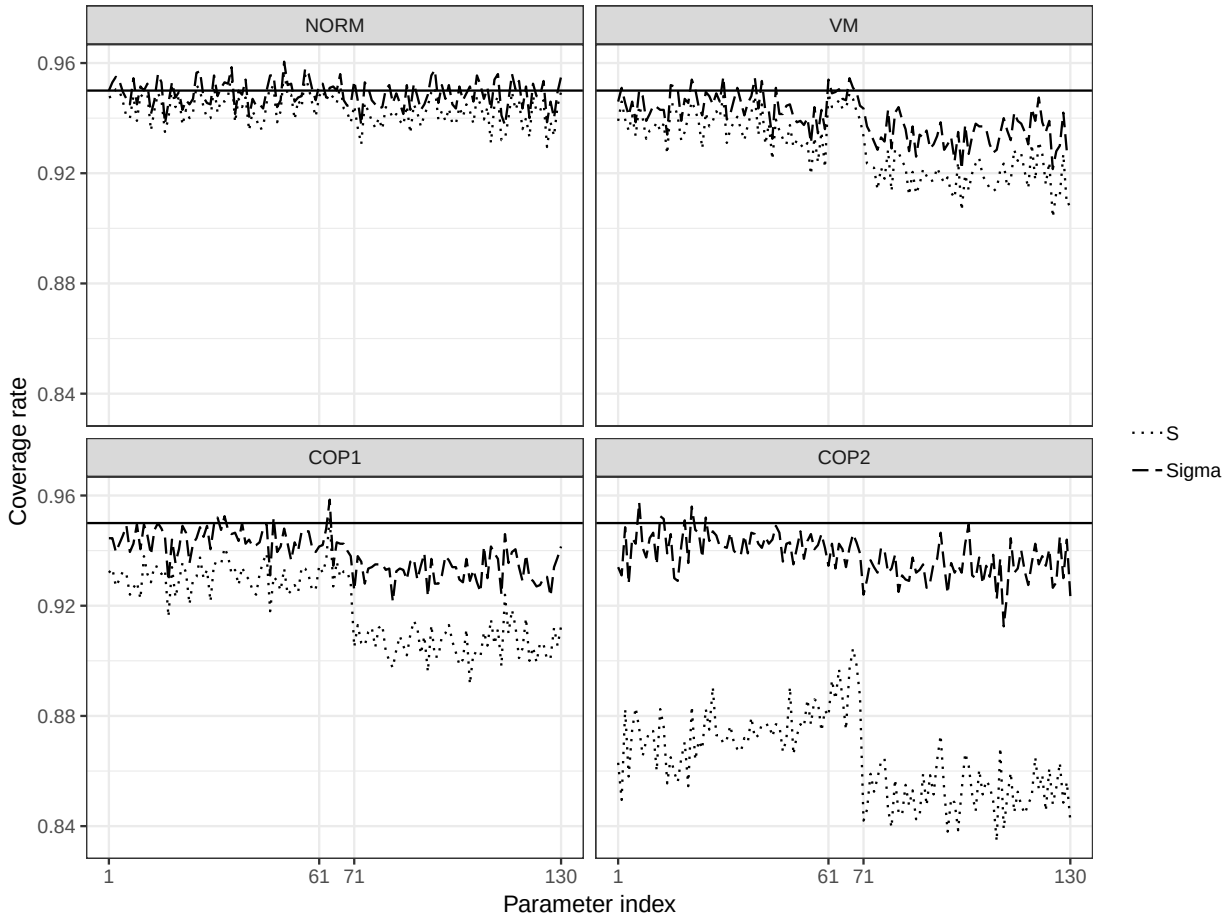

Figure 9. Confidence interval coverage rates for Model $\mathcal{M}_2$ at sample size $n = 2000$. NORM=Normal distribution. VM=Vale-Maurelli distribution. COP1= Gumbel copula distribution. COP2= Student's t copula distribution. S= Coverage rate based on SE(S). Sigma= Coverage rate based on SE(Σ).

Distribution	n	$T_1(\Sigma)$	$T_1(S)$	$T_2(\Sigma)$	$T_2(S)$	$T_3(\Sigma)$	$T_3(S)$
Normal	500	0.43	0.46	0.02	0.02	0.02	0.03
	2000	0.09	0.10	0.02	0.03	0.04	0.04
VM	500	0.39	0.76	0.00	0.03	0.01	0.11
	2000	0.09	0.15	0.01	0.02	0.04	0.07
COP1	500	0.59	1.00	0.00	0.24	0.01	0.65
	2000	0.11	0.31	0.00	0.02	0.04	0.13
COP2	500	0.01	1.00	0.00	1.00	0.00	1.00
	2000	0.00	1.00	0.00	0.27	0.00	0.99

Table 2

Model $\mathcal{M}_2$ rejection rates calculated at the $\alpha = 0.05$ level of significance. n= sample size. NORM=Normal distribution. VM=Vale-Maurelli distribution. COP1= Gumbel copula distribution. COP2= Student's t copula distribution. T_1=Satorra-Bentler test. T_2=Scaled-and-shifted test. T_3=Eigenvalue block averaging with 2 blocks. For each T_i results using both $\hat{W}_\Sigma$ and $\hat{W}_S$ are presented.

References

Asparouhov, T., & Muthen, B. (2010). Simple second order chi-square correction. *Mplus Technical Appendix. Retrieved from http://www. statmodel. com/download/WLSMV_new_chi21. pdf*.

Astivia, O. L. O., & Zumbo, B. D. (2018). On the solution multiplicity of the fleishman method and its impact in simulation studies. *British Journal of Mathematical and Statistical Psychology*.

Bedford, T., & Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 1031–1068.

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

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.

Costa, P. T., & McCrae, R. R. (1989). Neo pi/ffi manual supplement. *Odessa, FL: Psychological Assessment Resources*, *40*.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, *4*(1), 5.

Falk, C. F. (2018). Are robust standard errors the best approach for interval estimation with nonnormal data in structural equation modeling? *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 244-266. Retrieved from `https://doi.org/10.1080/10705511.2017.1367254` doi: 10.1080/10705511.2017.1367254

Foldnes, N., Foss, T., & Olsson, U. H. (2011). Residuals and the residual-based statistic for testing goodness of fit of structural equation models. *Journal of Educational and Behavioral Statistics*.

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

Foldnes, N., & Grønneberg, S. (2018). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling*, *25*, 101–114.

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.

Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate behavioral research*, *51*(2-3), 207–219.

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*, *82*(4), 1035–1051.

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 361-390. Retrieved from `https://doi.org/10.1080/10705510701301602` doi: 10.1080/10705510701301602

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 221–233).

Jöreskog, K., & Sörbom, D. (2015). *Lisrel 9.2.* Skokie, IL: Scientific Software.

Kojadinovic, I., & Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, *34*(9), 1–20. Retrieved from `http://www.jstatsoft.org/v34/i09/`

Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics*. Wiley.

Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating Nonnormal Multivariate Data Using Copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*(4), 547–565.

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation*

Modeling: A Multidisciplinary Journal, *24*(3), 383-394. Retrieved from https://doi.org/10.1080/10705511.2016.1269606 doi: 10.1080/10705511.2016.1269606

McCrae, R. R., & Costa, P. T., Jr. (2007). Brief versions of the neo-pi-3. *Journal of individual differences*, *28*(3), 116–128.

Muthén, L., & Muthén, B. (2010). *Mplus software (version 6.1).* Los Angeles, CA: Muthén & Muthén.

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

Sas/stat 14.1 user's guide [Computer software manual]. (2015). Cary, NC.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. V. Eye & C. Clogg (Eds.), *Latent variable analysis: applications for developmental research* (p. 399-419). Newbury Park, CA: Sage.

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223.

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)

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21-40. Retrieved from https://doi.org/10.1080/10705511.2017.1369088 doi: 10.1080/10705511.2017.1369088

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

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 471–494.

Xia, Y., Yung, Y.-F., & Zhang, W. (2016). Evaluating the selection of normal-theory weight matrices in the satorra–bentler correction of chi-square and standard errors. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 585–594.