

Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?

Stanislav Kolenikov*

Kenneth A. Bollen†

May 11, 2010

*Department of Statistics, University of Missouri, Columbia, MO 65211–6100, USA. Support from NSF grant SES-0617193 with funds from Social Security Administration is gratefully acknowledged. Corresponding author: kolenikovs@missouri.edu.

†Department of Sociology, University of North Carolina, Chapel Hill, 27599–3210 NC, USA. Support from NSF grant SES-0617276 is gratefully acknowledged.

Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?

Abstract

Heywood cases, or negative variance estimates, are a common occurrence in factor analysis and latent variable structural equation models. Though they have several potential causes, structural misspecification is among the most important. This paper explains how structural misspecification can lead to a Heywood case in the population, and provides several ways to test whether a negative error variance is a symptom of structural misspecification. We consider Wald tests based on a variety of standard errors, confidence intervals, bootstrap resampling, and likelihood ratio-type tests. In our discussion of Wald tests, we demonstrate which of the standard errors are consistent under different kinds of misspecification. We also introduce new tests based on the scaled chi-square difference: the test on the boundary and the signed root of the scaled chi-square difference. Our simulation study assesses the performance of these tests. We find that signed root tests and Wald tests based on the sandwich and the empirical bootstrap variance estimators perform best in detecting negative error variances. They outperform Wald tests based on the information matrix or distribution-robust variance estimators.

Keywords: bootstrap, factor analysis, Heywood case, improper solution, negative error variances, misspecified model, sandwich estimator, scaled chi-square difference, signed root, specification tests, structural equation models

1 Introduction

Negative variance estimates or “Heywood cases”¹ are common in factor analysis and structural equation models. Given the impossibility of these values in the population, researchers need to determine the reason for their occurrence. There is not a single cause of Heywood cases. Diagnostics must isolate their source. Among these causes are outliers (Bollen 1987), nonconvergence, underidentification (Van Driel 1978, Boomsma & Hoogland 2001), empirical underidentification (Rindskopf 1984), structurally misspecified models (Van Driel 1978, Dillon, Kumar & Mulani 1987, Sato 1987, Bollen 1989) or sampling fluctuations (Van Driel 1978, Boomsma 1983, Anderson & Gerbing 1984). There are outlier and influential case diagnostics for SEMs and factor analysis (Arbuckle 1997, Bollen 1987, Bollen & Arminger 1991, Cadigan 1995) and these provide a means to explore whether such cases are contributing to negative error variance estimates. Similarly, there are ways to detect nonconvergence, underidentification, and empirical underidentification. If none of these is the source of the improper solution, then sampling fluctuations or structural misspecification are the most likely remaining candidates. If after eliminating the other determinants a researcher can eliminate sampling fluctuations as the cause of improper solutions, then the evidence favors structural misspecification as the reason. And finding such would encourage the researcher to investigate possible flaws in the model specification. To do this, requires that we have tests of whether sampling error is the source of negative error variance estimates.

¹ The original paper (Heywood 1931) considers specific parameterizations of factor analytic models, in which some parameters used to generate correlation matrices were greater than 1. The use of the phrase “Heywood case” as referred to improper solutions in factor analysis can be traced back to early 1960s.

Van Driel (1978) suggested that researchers form confidence intervals from the maximum likelihood estimated asymptotic standard errors of the negative disturbance or error variance estimate and determine whether the interval includes zero. If it does, he suggested to interpret this as evidence that the population variance is positive but near zero, and that the negative estimate is due to chance. Similarly, Chen, Bollen, Paxton, Curran & Kirby (2001) recommend such confidence intervals along with z -tests and the likelihood ratio, Lagrange multiplier, and Wald tests where the latter apply to asymptotic tests of single or multiple negative error variance estimates. Chen et al. (2001), however, caution that some of these tests might not be fully accurate in samples with improper solutions. For instance, they found that the confidence intervals, z -tests, Wald tests, and Lagrange multiplier tests reject the null hypothesis too infrequently when the population error variance is set to zero. The likelihood ratio test rejects too frequently, but seems to perform the best (Chen et al. 2001, pp. 498–499).

One possibility to avoid Heywood cases is to restrict the range of estimates to be $[0, +\infty)$. There are however major problems with statistical inference, in particular, with regularity conditions for maximum likelihood requiring the true values to be in the interior of parameter space. If estimates are at the boundary (e.g., the variance of an error term is equal to zero), the estimates and statistical tests behave in unusual ways (Chernoff 1954, Andrews 1999, Andrews 2001). In SEM, the topic was raised by Shapiro (1985, 1988), but his results do not appear to be widely recognized, most likely due to their highly technical presentation. The interest in the topic has been renewed recently in Stoel, Garre, Dolan & van den Wittenboer (2006) and Savalei & Kolenikov (2008) who reviewed estimation with and without the inequality constraints such as requiring the parameters to have estimates in their “proper” ranges. Savalei & Kolenikov (2008) demonstrated that unconstrained estimation resulted in simpler asymptotic distributions and had greater power to detect structural misspecification related to negative error variance estimates.

Our paper has three major purposes that address testing negative error variances as a symptom of structural misspecification. First, we will explain the problems that can exist with the conventional test statistics in analyses with negative error variances. Second, we will recommend tests for Heywood cases that are asymptotically accurate even when the model is misspecified. Third, we will present a simulation experiment that looks at the robustness of the conventional test statistics and will examine the performance of the alternative tests that we propose.

The rest of the paper is organized as follows. The next section presents the statistical model and common estimators. A section on tests of negative variances follows with subsections on Wald tests, different types of standard errors that can be used to construct these tests, confidence intervals, bootstrap tests, likelihood ratio tests, and issues with multiple testing. After this, we give the results of our simulations in Section 4. Section 5 presents an empirical example showing implementation of the tests. Discussion of what we have learned in our analysis, what other approaches exist, and a summary of our recommendations concludes the paper.

2 Statistical model and estimators

Factor analysis and latent variable structural equation models are the two applications in which negative error variances are most discussed. Both of these models belong to the class of Structural Equation Models (SEMs).

2.1 Model and assumptions

In this section, we present the latent variable structural equation model using a modified version of Jöreskog's (1978) LISREL notation² that includes intercept terms. The *latent variable model* is:

$$\boldsymbol{\eta} = \boldsymbol{\alpha}_\eta + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (1)$$

The $\boldsymbol{\eta}$ vector is $m \times 1$ and contains the m latent endogenous variables. The intercept terms are in the $m \times 1$ vector of $\boldsymbol{\alpha}_\eta$. The $m \times m$ coefficient matrix \mathbf{B} gives the effect of the $\boldsymbol{\eta}$'s on each other. The n latent exogenous variables are in the $n \times 1$ vector $\boldsymbol{\xi}$. The $m \times n$ coefficient matrix $\boldsymbol{\Gamma}$ contains the coefficients for $\boldsymbol{\xi}$'s impact on the $\boldsymbol{\eta}$'s. An $m \times 1$ vector $\boldsymbol{\zeta}$ contains the disturbances for each latent endogenous variable. We assume that $\mathbb{E}(\boldsymbol{\zeta}) = \mathbf{0}$, $\text{COV}(\boldsymbol{\zeta}', \boldsymbol{\xi}) = \mathbf{0}$, and for now we assume that the disturbance for each equation is homoscedastic and uncorrelated *across cases* although the variances of $\boldsymbol{\zeta}$'s from different equations can differ, and these $\boldsymbol{\zeta}$'s can correlate across equations.³ The $m \times m$ covariance matrix $\boldsymbol{\Sigma}_\zeta$ has the variances of the ζ s down the main diagonal and the across equation covariances of the ζ s in the off-diagonal. The $n \times n$ covariance matrix of $\boldsymbol{\xi}$ is $\boldsymbol{\Sigma}_\xi$.

The *measurement model* is:

$$\mathbf{y} = \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (2)$$

$$\mathbf{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \quad (3)$$

The $p \times 1$ vector \mathbf{y} contains the indicators of the $\boldsymbol{\eta}$'s. The $p \times m$ coefficient matrix $\boldsymbol{\Lambda}_y$ (the ‘‘factor loadings’’) give the impact of the $\boldsymbol{\eta}$'s on the \mathbf{y} 's. The *unique factors* or ‘‘errors’’ are in the $p \times 1$ vector $\boldsymbol{\varepsilon}$. We assume that $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{COV}[\boldsymbol{\eta}\boldsymbol{\varepsilon}'] = \mathbf{0}$. The covariance matrix of $\boldsymbol{\varepsilon}$ is $\boldsymbol{\Sigma}_\varepsilon$. There are analogous definitions and assumptions for measurement equation (3) for the $q \times 1$ vector \mathbf{x} . We assume that $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$ are uncorrelated with $\boldsymbol{\xi}$ and in most models these disturbances and errors are assumed to be uncorrelated among themselves, though the latter assumption is not essential. We also assume that the errors are homoscedastic and uncorrelated across cases.

Researchers use equations (1), (2), and (3) to capture the relations that they hypothesize among the latent and observed variables by placing restrictions on the intercepts, coefficients, variances, or covariances. Each specified model implies a particular structure to the mean vector and covariance matrix of the observed variables:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E} \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_y + M(\boldsymbol{\alpha}_\eta + \boldsymbol{\Gamma}\boldsymbol{\mu}_\xi) \\ \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\mu}_\xi \end{pmatrix}, \quad (4)$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{COV} \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} M(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_\xi\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_\zeta)M' + \boldsymbol{\Sigma}_\varepsilon & M\boldsymbol{\Gamma}\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_\xi\boldsymbol{\Gamma}'M' & \boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_\xi\boldsymbol{\Lambda}_x' + \boldsymbol{\Sigma}_\delta \end{pmatrix}, \quad M = \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1} \quad (5)$$

where $\boldsymbol{\mu}$ is the mean vector of the subscripted variable, $\boldsymbol{\Sigma}$ is the covariance matrix of the subscripted variable, and the other symbols were previously defined.

² The notation differs in that intercept terms are represented by α 's and population covariance matrices are named with $\boldsymbol{\Sigma}$. Subscripts of these basic symbols make clear the variables to which they refer.

³ For the 2SLS estimator with heteroscedasticity in latent variable models see Bollen (1996b).

These assumptions are represented in the null hypothesis of:

$$\begin{aligned} H_0 : \Sigma &= \Sigma(\boldsymbol{\theta}) \\ \boldsymbol{\mu} &= \boldsymbol{\mu}(\boldsymbol{\theta}) \end{aligned} \quad (6)$$

where Σ is the population covariance matrix of the observed variables and $\boldsymbol{\mu}$ is the mean vector of the observed variables, $\mathbf{z}' = (\mathbf{y}', \mathbf{x}')$ is the vector of observed variables, $\boldsymbol{\theta}$ is a vector that contains all the intercepts, coefficients, variances, and covariances in the model to estimate, and $\Sigma(\boldsymbol{\theta})$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ are the model implied covariance matrix and implied mean vector that are functions of the parameters in $\boldsymbol{\theta}$. We assume that the parameters in $\boldsymbol{\theta}$ are identified (Bollen 1989, Ch. 8). A Heywood case is present when one or more elements of the main diagonals of $\widehat{\Sigma}_\zeta$, $\widehat{\Sigma}_\varepsilon$, $\widehat{\Sigma}_\delta$, or $\widehat{\Sigma}_\xi$ are negative.

2.2 Maximum likelihood (ML) estimator

The full information maximum likelihood (FIML) estimator, also shorthanded as ML, is the most widely used estimator of $\boldsymbol{\theta}$. The classic derivation of the FIML assumes that the observed variables come from multivariate normal distributions, though it maintains many of its desirable properties under less restrictive assumptions (Browne 1984, Satorra 1990). The maximum likelihood estimates are found by maximization of the log-likelihood

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \max l(\boldsymbol{\theta}, Z), \\ l(\boldsymbol{\theta}, Z) &= \sum_{i=1}^N \left[-\frac{p+q}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))' \Sigma^{-1}(\boldsymbol{\theta}) (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta})) \right] \\ &= -\frac{(p+q)N}{2} \ln 2\pi - \frac{N}{2} \ln |\Sigma(\boldsymbol{\theta})| - \frac{N}{2} \text{tr} \Sigma^{-1}(\boldsymbol{\theta}) S \end{aligned} \quad (7)$$

where S is the maximum likelihood estimate of covariance matrix, and the means $\boldsymbol{\mu}(\boldsymbol{\theta})$ are not modeled for covariance structure analysis. The SEM literature usually works with the equivalent problem of minimizing the goodness-of-fit criteria. The FIML objective function $F_{FIML}(\Sigma(\boldsymbol{\theta}), S)$ is the likelihood ratio test against an unstructured, or saturated, mean-and-covariance structure model normalized per observation:

$$\begin{aligned} F_{FIML}(\Sigma(\boldsymbol{\theta}), S) &= \ln |\Sigma(\boldsymbol{\theta})| + \text{tr}[\mathbf{S}\Sigma^{-1}(\boldsymbol{\theta})] + [\bar{\mathbf{z}} - \boldsymbol{\mu}(\boldsymbol{\theta})]' \Sigma^{-1}(\boldsymbol{\theta}) [\bar{\mathbf{z}} - \boldsymbol{\mu}(\boldsymbol{\theta})] \\ &\quad - \ln |\mathbf{S}| - (p+q) \end{aligned} \quad (8)$$

where S is the sample covariance matrix of the observed variables, $\bar{\mathbf{z}}$ is the vector of the observed variable sample means, and other symbols are already defined. The value of $\hat{\boldsymbol{\theta}}$ that minimizes $F_{FIML}(\Sigma(\boldsymbol{\theta}), S)$ or maximizes $l(\boldsymbol{\theta}, Z)$ is the FIML estimator.

2.3 Weighted least squares (WLS) estimator

Weighted Least Squares (WLS) is another popular estimator for SEMs. Beginning with Browne (1984), and further developed by Satorra (1990, 1992) and Satorra & Bentler (1990, 1994) we can write the WLS as the solution to the minimization problem of

$$F(\Sigma(\boldsymbol{\theta}), S, V_N) = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' V_N (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \quad (9)$$

leading to the estimating equations

$$\Delta V_N(\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) = 0 \quad (10)$$

Here $\mathbf{s} = \text{vech } S$, $\boldsymbol{\sigma}(\boldsymbol{\theta}) = \text{vech } \boldsymbol{\Sigma}(\boldsymbol{\theta})$ are the non-redundant elements of the two covariance matrices, vech is vectorization operator (Magnus & Neudecker 1999), and $\hat{\Delta} = \partial \boldsymbol{\sigma} / \partial \boldsymbol{\theta}$ is the derivative of the implied moments $\boldsymbol{\sigma}(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$. It has been shown (Lee & Jennrich 1979) that a special version of the iterative WLS minimization produces FIML estimates. Namely, if the weight matrix V_N

$$V_N^{(NT)} = \frac{1}{2} D'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) D, \quad (11)$$

called the normal theory weight matrix, is updated at each iteration, then the minimization procedure is numerically identical to Fisher scoring method of likelihood maximization. For a distribution that does not satisfy the asymptotic robustness conditions, asymptotically optimal (also called asymptotically distribution free) estimates (Browne 1984, Yuan & Bentler 1997, Yuan & Bentler 2007) are obtained by setting V_N to Ω_N , where Ω_N is an estimator of the asymptotic variance of the second moments:

$$\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \xrightarrow{d} N(\boldsymbol{\sigma}_*, \Omega_*) \quad (12)$$

Here, $\boldsymbol{\sigma}_*$ is $(p+q)(p+q+1)/2 \times 1$ vector of zeroes when the model structure is correctly specified, but is different from zero when structural misspecification is present. A popular choice of the estimator of Ω_* is equation (16.12) of Satorra & Bentler (1994), the empirical matrix of the fourth order moments of $\text{vech } S$:

$$\hat{\Omega}_N = \frac{1}{N-1} \sum_i (\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})'$$

$$\mathbf{b}_i = \text{vech}(y_i - \bar{y})(y_i - \bar{y})' \quad (13)$$

This estimator assumes that the model structure is correctly specified, so that $\boldsymbol{\sigma}_* = 0$.

Regardless of whether a researcher uses a FIML or a WLS estimator there are significance tests applicable to provide evidence as to whether a negative error variance is due to sampling fluctuations or to structural misspecifications. In the next section we present the main options for tests.

3 Tests of negative error variances

When testing for model misspecification the following null and alternative hypotheses are useful:

$$H_0 : \theta_k \geq 0$$

$$H_1 : \theta_k < 0 \quad (14)$$

where θ_k is the variance parameter of interest. If the null is rejected, then there is evidence that the model is structurally misspecified since a negative variance in the population is impossible. Note that H_0 here is a necessary, but not a sufficient condition for correct model specification. That is, if H_0 is rejected, then we have solid evidence that the model is structurally misspecified. If H_0 is true, then this is no guarantee of a correct specification.

We place the tests of (14) into three broad categories: (1) Wald tests and confidence intervals, (2) tests based on the bootstrap under the null hypothesis, and (3) likelihood ratio (LR) tests. We discuss these in the next few subsections. Our focus is on the most common situation of testing one parameter at a time. However, we cover multiple testing and simultaneous tests later in the paper.

The two main statistical characteristics of a test are its size and power. Size α is the probability of rejection when the null is true, or probability of type I error. Power of a test, $1 - \beta$, is the probability of rejection when the alternative is true. Tests can only be comparable in their power if their sizes are the same. Then a preferred test is the one that has a greater power. In the settings that we deal with, it is desirable for a test to have the correct size (5% or 10%, say), or at least achieve this size in large samples. If a test does not achieve the nominal size even asymptotically, it should not be used in practice, as its performance is essentially unknown. Also for the tests that have known sizes, the power is expected to increase as the sample size increases. For fixed alternatives, like the one investigated in our simulation studies, the power is expected to increase to 1 asymptotically.

Our preferences between the tests can be described by a lexicographic ordering:

1. The test must have the correct size (or at least the limitations of the tests regarding sample sizes, data distribution, and model specifications be known).
2. Among the tests of the same size, the most powerful tests are preferred.

3.1 Wald tests

When we test a single linear hypothesis $H_0 : \theta_k = 0$, the test statistic of the Wald test is

$$W = \frac{\hat{\theta}_k - \theta_{k0}}{\text{s.e.}[\hat{\theta}_k]} = \frac{\hat{\theta}_k}{\text{s.e.}[\hat{\theta}_k]} \quad (15)$$

and is printed by default in most software packages. Given $H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$, we are interested in one-sided tests of W in (15). That is, H_0 will be rejected only for

$$\frac{\hat{\theta}_k}{\text{s.e.}[\hat{\theta}_k]} < z_\alpha < 0 \quad (16)$$

where z_α is α -th quantile of the standard normal distribution, $\text{Prob}[z < z_\alpha] = \alpha$. For instance, the 5% test is to reject the null hypothesis if $\hat{\theta}_k/\text{s.e.}[\hat{\theta}_k] < -1.64$. Test of H_0 is a test of structural misspecification: its rejection suggests a structurally misspecified model. For the test statistic W to follow the standard normal distribution, requires that the asymptotic standard errors are consistent. There are several possible estimators of the variance-covariance matrix of parameter estimates that are the source of our asymptotic standard errors. Which asymptotic standard error is appropriate depends on the distribution from which the observed variables derive and whether the model is structurally misspecified. In the next series of subsections we present the asymptotic standard errors that are appropriate under different conditions.

3.1.1 Asymptotic standard errors for FIML

This section describes the standard errors that are applicable to distributions with no excess multivariate kurtosis. Assume that the model is correctly specified, so that $H_0 : \theta_k \geq 0$ is true. The standard likelihood

theory (van der Vaart 1998) provides that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is the inverse of the information matrix \mathcal{I} :

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}) \quad (17)$$

In turn, the latter must be estimated, either by the observed information matrix

$$\mathcal{I}_o = -\frac{1}{2} \frac{\partial^2 F_{FIML}(\boldsymbol{\Sigma}(\boldsymbol{\theta}), S)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (18)$$

or the expected information matrix,

$$\mathcal{I}_e = -\frac{1}{2} \mathbb{E} \frac{\partial^2 F_{FIML}(\boldsymbol{\Sigma}(\boldsymbol{\theta}), S)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (19)$$

evaluated at $\hat{\boldsymbol{\theta}}$. The choice of the observed vs. the expected information has received little attention in either statistical, econometric or SEM literature, although Savalei (forthcoming) pointed out that the expected information matrix-based tests perform poorly when the assumed model is incorrect. As this is a likely possibility, we use the observed information matrices in this paper. Researchers use square roots of the main diagonal entries of the estimated information matrix as the asymptotic standard errors for significance tests about individual parameters. In other words, these asymptotic standard errors would be substituted into equation (15) to enable a significance test of whether a negative error variance is statistically significant.⁴ Two assumptions are required for the validity of the asymptotic standard errors from (18) and (19) and hence the validity of the test statistic W from (15). The first is that the structural specification of the model is correct. In the context of negative variances this assumption is problematic in that under the alternative hypothesis ($H_1 : \theta_k < 0$) the model is structurally misspecified. The second assumption is that the observed variables come from distributions with no excess multivariate kurtosis. There are robustness conditions where these standard errors will still be accurate with excess kurtosis (Anderson & Amemiya 1988, Satorra 1990), but establishing that the robustness conditions hold might not be possible.

If either assumption fails, then the test of negative error variance might be inaccurate. This raises questions about the common practice of using the usual observed or expected information matrix based asymptotic standard errors to test the statistical significance of the negative variance estimates in a model.

3.1.2 Asymptotic standard errors for WLS

In this subsection we consider asymptotic standard errors when only the distributional assumption is violated.

As we discussed in section 2.3, the WLS estimator is the minimizer of $F(\boldsymbol{\Sigma}(\boldsymbol{\theta}), S, V_N) = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' V_N (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))$. This estimator applies even when the observed variables come from nonnormal distributions. The variance estimates for WLS estimator are obtained by the following delta-method argument (Browne 1984, Satorra & Bentler 1994). The estimates are obtained as solutions to the estimating equations (10):

$$\hat{\Delta} V_N (\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})) = 0.$$

⁴ In addition, the estimated asymptotic covariance matrix enables Wald tests on hypotheses about multiple parameters (Buse 1982, van der Vaart 1998) as we discuss below.

The population analogue of this equation is

$$\Delta_0 V(\boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta}_0)) = 0$$

where $\boldsymbol{\theta}_0$ is the parameter vector minimizing the fit criterion $F(\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\Sigma}, V)$ in the population, $\boldsymbol{\Sigma}$ is the true population covariance matrix, $\boldsymbol{\sigma}$ is its vectorization, V is the probability limit of the weight matrix V_N , and $\Delta_0 = \mathbb{E} \partial \boldsymbol{\sigma}(\boldsymbol{\theta}_0) / \partial \boldsymbol{\theta}$. Subtracting the two equations one from another, we obtain

$$0 = \hat{\Delta} V_N(\boldsymbol{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})) - \Delta_0 V(\boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta}_0)) = \hat{\Delta} V_N \boldsymbol{s} - \Delta_0 V \boldsymbol{\sigma} + \Delta_0 V \boldsymbol{\sigma}(\boldsymbol{\theta}_0) - \hat{\Delta} V_N \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$$

Approximating $\hat{\Delta} V_N$ by its probability limit $\Delta_0 V$, we obtain

$$\Delta_0 V(\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\sigma}(\boldsymbol{\theta}_0)) = \Delta_0 V(\boldsymbol{s} - \boldsymbol{\sigma})$$

Let us now take a first order expansion of $\boldsymbol{\sigma}(\boldsymbol{\theta})$ near $\boldsymbol{\theta}_0$:

$$\boldsymbol{\sigma}(\boldsymbol{\theta}) = \boldsymbol{\sigma}(\boldsymbol{\theta}_0) + \Delta_0'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|).$$

Hence, ignoring the last term,

$$\Delta_0 V \Delta_0'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \Delta_0 V(\boldsymbol{s} - \boldsymbol{\sigma}),$$

or

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (\Delta_0 V \Delta_0')^{-1} \Delta_0 V(\boldsymbol{s} - \boldsymbol{\sigma}).$$

Finally,

$$\begin{aligned} \text{As. V}[\boldsymbol{\theta}] &\approx \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'] \\ &= (\Delta_0 V \Delta_0')^{-1} \mathbb{E}[\Delta_0 V(\boldsymbol{s} - \boldsymbol{\sigma})(\boldsymbol{s} - \boldsymbol{\sigma})' V \Delta_0'] (\Delta_0 V \Delta_0')^{-1} \\ &= (\Delta_0 V \Delta_0')^{-1} \Delta_0 V \mathbb{E}[(\boldsymbol{s} - \boldsymbol{\sigma})(\boldsymbol{s} - \boldsymbol{\sigma})'] V \Delta_0' (\Delta_0 V \Delta_0')^{-1} \\ &= \frac{1}{N} (\Delta_0 V \Delta_0')^{-1} \Delta_0 V (\Omega_* + \boldsymbol{\sigma}_* \boldsymbol{\sigma}_*') V \Delta_0' (\Delta_0 V \Delta_0')^{-1} \end{aligned} \quad (20)$$

If the model is correctly specified, $\boldsymbol{\sigma}_* = 0$, resulting in expressions (2.12) of Browne (1984) or (16.9) of Satorra & Bentler (1994), except for a change in notation. The estimates are then obtained by replacing back Δ_0 with $\hat{\Delta}$, V with V_N , and Ω_* with an estimator such as (13):

$$\hat{\mathbb{V}}[\boldsymbol{\theta}] = \frac{1}{N} (\hat{\Delta} V_N \hat{\Delta}')^{-1} \hat{\Delta} V_N \hat{\Omega} V_N \hat{\Delta}' (\hat{\Delta} V_N \hat{\Delta}')^{-1} \quad (21)$$

The square root of the corresponding main diagonal element of equation (21) gives an asymptotic standard error that we use in equation (15) when the observed variables come from distributions with excess multivariate kurtosis. These asymptotic standard errors protect us from inaccurate standard errors due to distributional assumptions violations, but they do not protect us from structurally misspecified models.

Indeed, if the model is misspecified, the estimator (13) underestimates the internal part of (20) by $\boldsymbol{\sigma}_* \boldsymbol{\sigma}_*'$. As a result, the standard errors are underestimated, the confidence intervals are too narrow, and the Wald tests based on these standard errors reject too often. Given that structural misspecification occurs when we have a negative error variance under the alternative hypothesis ($H_1 : \theta_k < 0$), this limitation of WLS asymptotic standard errors is a real concern. To obtain accurate standard errors, $\boldsymbol{\sigma}_*$ needs to be estimated (as suggested by Yuan & Hayashi (2006)), or a different approach to variance estimation may be taken.

3.1.3 Huber asymptotic standard errors

In this section, we discuss another standard error estimate that is robust not only to distributional violations, but to structural misspecifications as well.

The WLS standard errors, often referred to as “robust” standard errors, have been extensively used in applied work in the last 15–20 years to correct for violations of excess multivariate kurtosis. However, little work was done on the robustness of the FIML estimates and associated standard errors (17) or (21) when the structure of the model, as opposed to the distribution of the observed variables, is not specified correctly.

The earliest and the most general work on misspecified models dates back to Huber (1967) who showed consistency and asymptotic normality of the quasi-MLEs when the actual density of the data is different from the assumed model. Similar treatment was given in econometrics by White (1982) who used somewhat milder (and easier to check) regularity conditions on the smoothness of the objective function. The specification of the multivariate normal structural equation model given by (5) and (7) might be incorrect in that the distribution of the data is not multivariate normal, or that an incorrect covariance structure is being fit to data.

Suppose the i.i.d. data $Z_i, i = 1, \dots, N$ come from an unknown distribution, and the estimates are derived as the solutions of estimating equations

$$\Psi(\boldsymbol{\theta}; \mathbf{Z}) \equiv \frac{1}{N} \sum_{i=1}^N \psi(\boldsymbol{\theta}; Z_i) = 0 \quad (22)$$

where $\psi(\boldsymbol{\theta}, Z_i)$ is the vector of contributions from observation i . Often, the vector of estimating equations $\Psi(\cdot)$ in (22) is the gradient of an objective function

$$Q(\boldsymbol{\theta}; \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^N q(\boldsymbol{\theta}; Z_i) \rightarrow \max_{\boldsymbol{\theta}} \quad (23)$$

where $q(\cdot)$ are observation level contributions. These expressions apply to both the FIML and the WLS estimation methods presented above. The likelihood (7) is indeed the sum of individual level terms

$$q_{ML}(\boldsymbol{\theta}; Z_i) = -\frac{p+q}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

Estimating equations from WLS (10) are another example, in which case the objective function is comprised of

$$q_{WLS}(\boldsymbol{\theta}; Z_i) = (\text{vech}[(z_i - \bar{z})(z_i - \bar{z})'] - \boldsymbol{\sigma}(\boldsymbol{\theta}))' V_N (\text{vech}[(z_i - \bar{z})(z_i - \bar{z})'] - \boldsymbol{\sigma}(\boldsymbol{\theta}))$$

assuming that only the covariance structure is modeled. For this objective function, the estimation equations are given by

$$\psi(\boldsymbol{\theta}; Z_i) = -2(\text{vech}[(z_i - \bar{z})(z_i - \bar{z})'] - \boldsymbol{\sigma}(\boldsymbol{\theta}))' V_N \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\theta}}$$

Estimating equations can also be obtained from other considerations, such as the moment conditions for model-implied instrumental variable estimation methods in Bollen (1996a, 1996b).

The estimates obtained from (22) or (23) are referred to as M -estimates, after Huber (1974). Under certain regularity conditions given in the aforementioned papers,⁵ the estimators $\hat{\boldsymbol{\theta}}_N$ obtained by solving (22) or maximizing (23) exist, are unique, consistent, and asymptotically normal:

$$\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) &\xrightarrow{d} N(0, A^{-1}BA^{-T}), \\ A &= \mathbb{E} D\Psi(\boldsymbol{\theta}_0, Z), \quad B = \mathbb{E} \Psi(\boldsymbol{\theta}_0, Z)\Psi(\boldsymbol{\theta}_0, Z)^T \end{aligned} \quad (24)$$

where $D\Psi(\boldsymbol{\theta}, Z)$ is the matrix of derivatives, $(D\Psi(\boldsymbol{\theta}, Z))_{ij} = \frac{\partial}{\partial \theta_j} \Psi_i(\boldsymbol{\theta}, Z)$, $\boldsymbol{\theta}_0$ is the value that solves the population equivalent of (22) or (23), and A^{-T} is the transposed inverse of the matrix A . In the FIML problem, A is the matrix of second derivatives, and B is the matrix of the outer product of gradients. The expression for the variance of the estimator is known as the information sandwich. The expression is very general, and special forms of it have been derived in the structural equation modeling literature (Browne 1984, Arminger & Schoenberg 1989, Satorra 1990, Satorra & Bentler 1994, Yuan & Hayashi 2006), econometrics (Eicker 1967, White 1980, West & Newey 1987), survey statistics (Binder 1983, Skinner 1989), and in other contexts. For a review of the history and applications of the sandwich estimator, see Hardin (2003). The equation (24) is obtained by the multivariate delta-method (Amemiya 1985, Ferguson 1996, van der Vaart 1998), and the components of the typical delta-method formula are easily seen: the matrix B gives the variance of the “original” asymptotically normal expressions (22), and the matrix A is the derivative of the transformation.

In the context of M -estimation, parameters are treated formally as unknowns in the nonlinear model. Interpretation of the parameters is left to the researcher, who might view them as variances, regression slopes, etc., but there is nothing in the mathematical formulation of the maximization problem, and the way the problem is set up, that binds the researcher for particular ranges of values for those parameters. Negative variances, though not making substantive sense, would still constitute valid parameter estimates that optimize a certain objective function (in the sample or in the population), and valid (Wald type) tests can be constructed for them based on the asymptotic covariance matrix given by (24). Note that while in properly structurally specified models all consistent methods converge to the same population values of the parameters, the situation is more complicated in misspecified models: different objective functions such as FIML or WLS may give rise to different values of $\boldsymbol{\theta}$ that minimize their respective objective functions in the population (Yuan & Chan 2005). Thus different estimation methods might converge to different values. We confine our attention to FIML estimator in this paper.

The arguments leading to (24) are asymptotic in their nature, and it has been shown that the finite sample performance of the sandwich estimator may be disappointing. Carroll, Wang, Simpson, Stromberg & Ruppert (1998) and Kauermann & Carroll (2001) argue that this small sample performance is the price one has to pay for its consistency, as it is typical in robust statistics to trade off efficiency for robustness. They show that the estimator is more variable than the naïve variance estimator (the inverse of the Fisher

⁵ The regularity conditions include measurability of the objective function over the space of X and sufficient smoothness with respect to $\boldsymbol{\theta}$, boundedness of the objective function, local compactness of the space of $\boldsymbol{\theta}$, uniqueness and sufficient separation of population maximum at the interior of the parameter space. The sets of regularity conditions might vary; e.g., Huber (1967) provides minimal conditions that do not require differentiability of the $\Psi(\cdot)$; on the other hand, White (1982) analyzes the misspecification problems under somewhat simpler conditions that involve differentiable functions, and shows sufficiency of his conditions for those of Huber (1967).

information matrix). As a result, coverage of the confidence intervals based on this estimator is below the nominal levels. They also propose corrections improving the performance of the estimator in the linear regression context.

The asymptotic variance (24) requires the first derivatives of the estimating equations, or the first and the second derivatives of the objective function. The necessary derivatives can be obtained analytically (Neudecker & Satorra 1991, Yuan & Hayashi 2006) or by numeric differentiation (Jennrich 2008). While the aforementioned papers provide the analytical expressions in their utmost generality, no SEM software currently supports these standard errors. The necessary components, however, are easily available as a by-product of gradient-based optimization procedures in other general purpose statistical packages such as Stata or R, implemented as `_robust` command and `sandwich` package, respectively. The Appendix provides some details of such implementation, and reviews the associated numerical issues. In our approach, and also in the GLLAMM approach (Rabe-Hesketh, Skrondal & Pickles 2004, Skrondal & Rabe-Hesketh 2004, Rabe-Hesketh & Skrondal 2005), the expressions in (24) are estimated directly, with the expectations replaced by sample means of the estimating equation derivatives and outer products that are automatically produced by Stata software during optimization steps. As the theoretical expectations are replaced by the empirical ones, we shall refer to this estimator as *empirical sandwich* estimator.

A colloquial term “robust” is often applied to various versions of the sandwich estimator, but in fact this term is rather unfortunate. First, the resulting estimator is not robust in terms of technical definitions of robustness (Huber 1974). Its components are linear in the likelihood scores and second derivatives, and thus their influence functions are unbounded. They can be unstable in the presence of multivariate outliers. For an example of implementation of robust procedures in structural equation modeling, see Moustaki & Victoria-Feser (2006). Second, in finite samples, the sandwich estimator may not lead to very good estimates of the standard errors. Third, different versions of the sandwich estimator deal with different violations of the original model assumptions. In the linear regression context where the sandwich-type estimators are best studied, the heteroskedasticity-robust variance estimator of Eicker (1967) and White (1980) will be inconsistent under serial or cluster correlations of observations. In turn, either misspecifications can be corrected (West & Newey 1987, White 1996). Even less fortunate is the use of the term “robust standard errors” when applied to the WLS standard errors (21), as they are inconsistent under structural misspecification.

The square roots of the diagonal entries of the asymptotic covariance matrix given in (24) give us the asymptotic standard errors based on the Huber sandwich estimator. When substituted into equation (24), we have a large sample test statistic that is robust to both distributional and structural misspecification problems. In other words, we have a means to test $H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$ that permits both nonnormal data and incorrect model specification. However, these are large sample results. Our simulations help to evaluate their performance across sample sizes typical in empirical research.

The proposed procedure, and its numeric implementation in particular, is identical to the infinitesimal jackknife procedure of Jennrich (2008). The latter computes pseudovalues of parameters for each observation in the data set, and provides the variance-covariance matrix of parameter estimates as the variance-covariance matrix of the pseudovalues. The pseudovalues, in turn, are infinitesimal changes in the estimating equations associated with the changes in the parameters, i.e., the derivatives that appear in (23) and (24).

3.1.4 The bootstrap standard errors

Yet another approach to estimation of the standard errors is based on resampling from the distribution of data. Popular resampling approaches include the jackknife (Shao & Tu 1995) and the bootstrap (Efron 1979, Efron & Tibshirani 1994). Since the latter method offers a richer set of inference techniques, we shall concentrate on it. If the sample X_1, \dots, X_N is obtained from distribution \mathcal{F} , and the quantity of interest is the sample statistic $T_N = T(X_1, \dots, X_N)$, then the basic form of the bootstrap assesses variability of T_N by considering the distribution of $T_*^{(b)}$ over possible samples $X_{*1}^{(b)}, \dots, X_{*N}^{(b)}$ where for b -th sample, $X_{*i}^{(b)}$ are sampled with replacement from the empirical distribution \mathcal{F}_N of the data X_1, \dots, X_N . In the exact bootstrap, all possible subsamples are considered. In practical situations, this would be computationally infeasible, since the number of possible samples N^N is combinatorially large, and a sufficiently large number of subsamples B is taken instead. Another potential complication for SEM is that some of the bootstrap samples produce singular sample covariance matrices for which the SEM analysis cannot be conducted. Such samples are usually discarded in practical application of the bootstrap. In large samples, the empirical distribution \mathcal{F}_N will be “close” to the true distribution \mathcal{F} , and thus samples from \mathcal{F}_N will provide the estimate of the distribution of T_N “close” to the true distribution. The rigorous theory of the bootstrap (Hall 1992, Shao & Tu 1995) specifies what exactly “close” should mean in the above statements, as well as gives examples when the bootstrap fails to give correct answers (Canty, Davison, Hinkley & Ventura 2006).

The bootstrap estimate of the variance $[v_B(\hat{\theta})]$ of the sample estimates is given by

$$v_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_*^{(b)} - \bar{\theta}_*)^2 \quad (25)$$

so for the bootstrap standard error $[s.e._B(\hat{\theta})]$ we have

$$s.e._B(\hat{\theta}) = \sqrt{\frac{1}{B} \sum_{b=1}^B (\hat{\theta}_*^{(b)} - \bar{\theta}_*)^2}$$

where $\hat{\theta}_*^{(b)}$ is the estimate obtained from b -th bootstrap sample and $\bar{\theta}_*$ is the mean $\left(= \frac{\sum_{b=1}^B \hat{\theta}_*^{(b)}}{B} \right)$ of the bootstrap estimates. The number of replications B is the tuning parameter of the bootstrap procedure (Efron & Tibshirani 1994, Sec. 6.4). Since low values of B lead to large Monte Carlo variability, or instability, of the standard errors (3.1.4), the typical values of B usually range from 100 to 1000.

The bootstrap standard error $[s.e._B(\hat{\theta})]$ provides an asymptotic standard error that we can use in equation (15) for testing the statistical significance of the negative error variance. This asymptotic standard error is similar to the Huber standard error in that it permits excess multivariate kurtosis and it is not undermined by structural misspecifications. However, like the Huber standard errors its properties are asymptotic and we do not know its performance in finite samples.

A variation of the bootstrap scheme popular in structural equation modeling is the bootstrap from the null distribution. This procedure is especially helpful to provide inference for the overall goodness of fit test (32) when the data are non-normal. To ensure that the data conform to the null hypothesis, the bootstrap samples are taken from the distribution of

$$\tilde{z}_i = \Sigma(\theta)^{1/2} S^{-1/2} z_i. \quad (26)$$

Table 1: Variance estimators in distributionally and structurally misspecified models.

Distributional specification	Structure specification	
	Correct	Incorrect
Correct	IM, WLS, ES, EB, BSB	ES, EB
Incorrect	WLS, ES, EB, BSB	ES, EB

Analytic standard errors: IM, observed or expected information matrix;
WLS, WLS standard errors; ES, empirical sandwich.
Resampling standard errors: EB, empirical bootstrap;
BSB: Bollen-Stine bootstrap.

Beran & Srivastava (1985) proposed a general bootstrap testing framework for covariance matrices using this transformation, and Bollen & Stine (1992) explained it in the context of SEMs. Here, $\Sigma(\theta)$ is any covariance matrix that satisfies the researcher’s model. It is often taken to be the implied moment matrix based on estimation results, $\Sigma(\hat{\theta})$, so as to minimize the disturbance of the distribution of z . By employing this transformation, the researcher obtains an auxiliary data set for which the model is true, yet the multivariate aspects of the data such as kurtosis are as close to those of the original data as possible.

The distinguishing feature of Bollen–Stine bootstrap is that it forces the model to be true in the bootstrap population. While this produces the correct null distribution of the test statistic (32), it is unclear whether the resulting variance estimates based on (25) are consistent. Yuan & Hayashi (2006) argue that the Bollen–Stine bootstrap is asymptotically equivalent to WLS standard errors (21). Since the latter are biased under structural misspecification, Bollen–Stine bootstrap standard errors will also be biased.

3.2 Comparison of asymptotic standard errors

Table 1 is a convenient summary of the different asymptotic standard errors available to use in the Wald test (15). The choice of which standard error to use primarily depends on whether there is excess kurtosis or structural misspecification. With correct structural specification and no excess kurtosis, all the asymptotic standard errors are consistent estimators of variability that could be used in equation (15). If the model is true and the only assumption violation is excess multivariate kurtosis (or, to be precise, the conditions of asymptotic robustness of FIML (Anderson & Amemiya 1988, Satorra 1990, Satorra 1992) are not satisfied), then the information matrix based FIML asymptotic standard errors are likely inaccurate, but all other asymptotic standard errors remain consistent estimators. The upper right cell of Table 1 corresponds to having no excess multivariate kurtosis, but a structurally misspecified model. In this situation, the only appropriate asymptotic standard errors are the Huber sandwich and empirical bootstrap standard errors. These same two asymptotic standard errors are consistent estimators when there are structural misspecifications and excess kurtosis. Thus we see that the asymptotic standard errors that require the fewest assumptions are the Huber sandwich and the empirical bootstrap asymptotic standard errors in that they apply to a structurally misspec-

ified model with observed variables from a distribution with excess multivariate kurtosis. It is yet unclear whether the empirical bootstrap or the Huber sandwich standard errors would be preferable in a given finite sample situation. While the bootstrap is often viewed as a finite-sample method, its justifications are still asymptotic. Our simulations (Section 4) provide some evidence on these issues.

3.3 Confidence intervals

A second way of checking $H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$ and hence to check structural misspecification is to form a confidence interval (CI) around the variance estimate and to see whether the upper limit includes nonnegative values. There are two common ways to construct confidence intervals: by using asymptotic normality and the asymptotic standard errors (analytical approach), or by using the bootstrap (resampling approach).

In general, the analytical two-sided confidence interval is

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \times \text{s.e.}(\hat{\theta}) \quad (27)$$

where $\hat{\theta}$ is the parameter estimate, $\text{s.e.}(\hat{\theta})$ is the asymptotic standard error of $\hat{\theta}$, and $z_{1-\frac{\alpha}{2}}$ is the corresponding percentile of the standard normal distribution determined by the Type I error probability. Analytical construction of the confidence intervals leads to the inference outcomes identical to those of Wald tests for a single parameter. If the true value is outside the confidence interval, it also means that the z -score of the Wald test exceeds the critical value, and so does the χ_1^2 Wald test statistic. The confidence intervals and Wald tests are parallel to one another.

Since we are interested in one-sided hypotheses tests (14), we would also be interested in one-sided analytical confidence intervals

$$(-\infty, \hat{\theta} + z_{1-\alpha} \times \text{s.e.}(\hat{\theta})) \quad (28)$$

analogous to the one-sided hypothesis testing problem. Again, rejection of the one-sided test is identical to lack of coverage of the true value by the one-sided CI.

As was true for the Wald test (15), there are several choices for the asymptotic standard errors. Which one is appropriate depends on whether the distributional assumptions are satisfied for the observed variables and whether the model is correctly specified. The discussion on asymptotic standard errors from the section on Wald tests carries over to the confidence intervals and we will not repeat it here except to say that if the assumptions that underlie the asymptotic standard errors do not hold, then the accuracy of the standard errors and the confidence intervals that depend on them cannot be guaranteed.

An alternative approach to construction of the confidence intervals is based on the bootstrap. Using the bootstrap framework of section 3.1.4, the two-sided bootstrap percentile confidence interval of level $1 - \alpha$ is

$$(\hat{\theta}_*^{[B\alpha/2]}, \hat{\theta}_*^{[B(1-\alpha/2)]}) \quad (29)$$

where $\hat{\theta}_*^{[1]} \leq \hat{\theta}_*^{[2]} \leq \dots \leq \hat{\theta}_*^{[B]}$ are ordered bootstrap realizations of the estimate $\hat{\theta}$. A one-sided confidence interval of level α is

$$(-\infty, \hat{\theta}_*^{[B(1-\alpha)]}). \quad (30)$$

Regardless of how a confidence interval was constructed, our CI-based test of the Heywood case consists of verifying whether the confidence interval covers zero. If it does not, it should be viewed as evidence against the null hypothesis of correct structural specification.

3.3.1 The Bollen-Stine bootstrap test

Another way to control for lack of multivariate normality and asymptotic robustness is to utilize the Bollen-Stine rotating bootstrap. To ensure that the distribution corresponds to the null hypothesis, the matrix $\Sigma(\boldsymbol{\theta})$ must come from a model in which the relevant Heywood case is eliminated by replacing $\theta_k = 0$. Other estimates necessary to construct $\Sigma(\boldsymbol{\theta})$ are obtained by maximization subject to this constraint. The p -value of the Bollen-Stine bootstrap test is the proportion of times the actual estimate $\hat{\theta}_k$ is below the estimated $\hat{\theta}_{k*}^{(b)}$ in the b -th bootstrap sample:

$$p = \frac{1}{B} \sum_b \mathbb{I}[\hat{\theta}_k < \hat{\theta}_{k*}^{(b)}] \quad (31)$$

3.4 Likelihood ratio tests

Likelihood ratio (LR) tests or “chi-square” tests are another tool for testing statistical significance in SEMs estimated with FIML. The most widely used LR test in SEM is of $H_0 : \Sigma = \Sigma(\boldsymbol{\theta}) \ \& \ \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$. This is a test of whether the overidentification restrictions implied by the whole model hold. The moment structure hypothesis in (6) is tested by forming

$$T = NF_{FIML}[\hat{\boldsymbol{\theta}}] \quad (32)$$

where N is the sample size,⁶ and then under the null hypothesis T asymptotically follows a χ^2 distribution with degrees of freedom equal to $r = \frac{1}{2}(p+q)(p+q+3) - t$ where t is the number of distinct parameters in $\boldsymbol{\theta}$.

In addition, when one model is nested within another by imposing some equality restrictions (leading to two-sided tests; see Sec. 3.4.3), the difference in the test statistics for the two nested models forms an asymptotic chi square statistic to test whether the most restricted of the two model fits as well as the less restrictive one. If the model is estimated with and without restrictions imposed by the null hypothesis, then the LRT statistic T is twice the difference of the log-likelihoods (7), or N times the difference of the minimized values of the objective function (8):

$$T = -2(l(\hat{\boldsymbol{\theta}}_0, Z) - l(\hat{\boldsymbol{\theta}}_1, Z)) = N[F_{FIML}(\Sigma(\boldsymbol{\theta}_1), S) - F_{FIML}(\Sigma(\boldsymbol{\theta}_0), S)],$$

$$\hat{\boldsymbol{\theta}}_0 = \arg \max_{\boldsymbol{\theta} \in \Theta_0} l(\boldsymbol{\theta}, Z), \quad \hat{\boldsymbol{\theta}}_1 = \arg \max_{\boldsymbol{\theta} \in \Theta_1 \cup \Theta_0} l(\boldsymbol{\theta}, Z) \quad (33)$$

3.4.1 Scaled and adjusted tests

When the observed variables come from a distribution with excess kurtosis and the asymptotic robustness conditions are not met, the (quasi-)likelihood ratio test statistic (32) loses its pivotal χ^2 form and becomes a

⁶ Sometimes, $N - 1$ is used as a multiplier, which is asymptotically equivalent to (32).

sum of weighted χ^2 :

$$T \xrightarrow{d} \sum_{j=1}^r \alpha_j X_j, \quad X_j \sim \text{i.i.d.} \chi_1^2. \quad (34)$$

Scalars α_j are eigenvalues of the matrix $U_0\Omega$ with

$$U_0 = V - V\Delta_0(\Delta_0'V\Delta_0)^{-1}\Delta_0'V, \quad (35)$$

where Δ_0 and V were defined in section 3.1.2. Satorra & Bentler (1994) proposed to use the scaled statistic

$$T_{\text{sc}} = \frac{T}{\hat{c}}, \quad \hat{c} = \frac{1}{r} \text{tr}[\hat{U}_0\hat{\Omega}_N] \quad (36)$$

referred to χ_r^2 , and adjusted statistic

$$T_{\text{adj}} = \frac{\hat{d}}{\hat{c}}T, \quad \hat{d} = \frac{(\text{tr}[\hat{U}_0\hat{\Omega}])^2}{\text{tr}[(\hat{U}_0\hat{\Omega}_N)^2]} \quad (37)$$

referred to $\chi_{\hat{d}}^2$ where the degrees of freedom \hat{d} might be a non-integer number. Here \hat{U}_0 is (35) evaluated at $\hat{\theta}$. These are standard tests implemented in most SEM software packages.

We can also implement the scaling version Satorra-Bentler corrections (36) to test nested models (Satorra & Bentler 2001). If T_1 and T_2 are the χ^2 goodness of fit statistics for two nested models, r_1 and r_2 are respective degrees of freedom, and \hat{c}_1 and \hat{c}_2 are Satorra-Bentler scaling corrections, then Satorra-Bentler scaled difference test can be computed as

$$\begin{aligned} \bar{T}_d &= \frac{T_1 - T_2}{\hat{c}_d}, \\ \hat{c}_d &= \frac{r_1\hat{c}_1 - r_2\hat{c}_2}{r_1 - r_2} \end{aligned} \quad (38)$$

referred to χ^2 with $r_1 - r_2$ degrees of freedom. Again, this test is relevant for the two-sided situation (41).

3.4.2 Tests on the boundary of the parameter space

To apply the standard or corrected chi square tests to our hypotheses about negative variances introduces complications. Recall that our null and alternative hypotheses are $H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$, that is, a one-sided test. For one-sided testing (14), the asymptotic distribution of the usual likelihood ratio test depends on the true value of the parameter. We distinguish two cases.

In the generic null situation when the true value of θ_k is *strictly* greater than zero, its (consistent) unrestricted estimate $\hat{\theta}_{k1}$ will also tend to be positive in large samples. By the law of large numbers, $\text{Prob}[\hat{\theta}_{k1} > 0] \rightarrow 1$ as $N \rightarrow \infty$. If $\hat{\theta}_{k1} > 0$, the estimates with and without the constraints imposed by the null hypothesis coincide, thus giving $T^b = 0$. (Superscript b stands for the effect of the boundary of the parameter space.) Negative values, on the other hand, produce non-zero values of the statistic T^b , but the probability of negative values decreases to 0 as the sample size increases to infinity (Savalei & Kolenikov 2008).

In the special case, still under the null hypothesis, that $\theta_k = 0$ (i.e., we have found a perfect indicator of a latent variable, which is questionable), $\hat{\theta}_{k1}$ will take positive and negative values approximately half of the time each. While the positive values produce $T_+^b = 0$, the negative values produce the test statistic $T_-^b = (\hat{\theta}_{k1}/\text{s.e.}[\hat{\theta}_{k1}])^2$. The latter has the distribution function $\text{Prob}[T_-^b < t] = \text{Prob}[Z^2 < t | Z < 0] = \text{Prob}[Z^2 < t] = \text{Prob}[\chi_1^2 < t]$, where Z is a standard normal variate, and the conditioning in the second equality is irrelevant since the distribution of Z^2 is the same for both positive and negative values of Z . Overall, the asymptotic distribution is given by a mixture (Chernoff 1954, Perlman 1969, Shapiro 1985, Andrews 1999, Andrews 2001, Savalei & Kolenikov 2008),

$$\text{Prob}[T^b \leq c] \rightarrow \begin{cases} 0, & c < 0, \\ \frac{1}{2} + \frac{1}{2} \text{Prob}(\chi_1^2 < c), & c \geq 0 \end{cases} \quad (39)$$

We shall denote this asymptotic distribution by $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$; it is also known as $\bar{\chi}$ distribution (Kûdo 1963). This is the most conservative distribution across all possible null hypotheses, most favorable for any alternative. Only with this distribution can a test with a non-trivial asymptotic size different from 0 or 1 be developed.

Note that distribution (39) of the likelihood ratio test statistic will only work with a single parameter. Furthermore, the asymptotic distribution for the test statistic, T^b , is only applicable when the distributional assumption for the observed variables is met (i.e., there is no excess kurtosis) and that the structural specification is correct. To account for the effect of the boundary in the situations where asymptotic robustness conditions are violated, we can consider an appropriate modification of the scaled difference test (38). If the variance estimate in question is positive, the test statistic \bar{T}_d^b is zero, while if the estimate is negative, it is computed according to (38). The overall distribution of the test is approximately

$$\text{Prob}[\bar{T}_d^b \leq c] \approx \begin{cases} 0, & c < 0, \\ \frac{1}{2} + \frac{1}{2} \text{Prob}(\chi_1^2 < c), & c \geq 0 \end{cases} \quad (40)$$

The quality of the approximation depends on whether the two-sided scaled difference statistic (38) is well approximated by the χ^2 distribution with one degree of freedom. If it is, then the approximation (40) improves with the sample size.

Note that tests on the boundary are still intended to test the null situation (14) of correct specification. The test (39) will work under no excess multivariate kurtosis, and test (40), under arbitrary kurtosis, but they both assume the model is correctly specified. If the model is structurally misspecified, the distributions of these tests will be different from the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ mixture. Savalei & Kolenikov (2008) discuss in detail the asymptotic and finite sample approximations that this test utilizes.

3.4.3 Two-sided tests

The last subsection explained the difficulty with testing the one sided hypothesis of $H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$. Different null and alternative hypotheses tied to tests of negative variances are somewhat simpler to handle:

$$\begin{aligned} H_0 : \theta_k &= 0 \\ H_1 : \theta_k &\neq 0 \end{aligned} \quad (41)$$

Here the null hypothesis is that the variance is zero. This is the smallest plausible value consistent with a model being correctly specified. If θ_k were an indicator error variance, then H_0 implies that there is no measurement error in the indicator; if θ_k were the error variance of an equation, then H_0 implies that there is no error in predicting the outcome; and finally if θ_k were the variance of a latent variable, then H_0 implies that the latent variable does not exist. The alternative hypothesis is not directional, though in practice the test would likely be applied when the sample variance estimate is negative and the researcher wants to determine whether it is within sampling error of zero. However, there are occasions where even a positive variance estimate might result in a test of $H_0 : \theta_k = 0$. For instance, a researcher might want to test for the presence of a latent variable and the variance estimate of the variable is positive but small. Testing $H_0 : \theta_k = 0$ could provide evidence as to the existence of the latent variable. More generally, testing $H_0 : \theta_k = 0$ is a different setup that could provide evidence relevant to negative variances even though the alternative hypothesis is nondirectional. The two-sided LR test is more straightforward than the one-sided tests described in the previous subsection. A nested chi square difference test is available where the model is first estimated without any constraints on the variances (unrestricted model) and then estimated when setting the problematic variance to zero. An LR test that compares the restricted and unrestricted form is a test of $H_0 : \theta_k = 0$. Rejection means a significance difference of the variance from zero. If the sample estimate is negative it suggests a structurally misspecified model.

3.4.4 Signed root tests

As noted in the discussion of expression (14), our interest lies in one-sided tests of Heywood cases, while the tests just discussed are two-sided tests of $H_0 : \theta_k = 0$ vs. $H_1 : \theta_k \neq 0$. There is a class of tests that addresses the desire for one-sided alternative hypotheses known as signed root tests. For those tests, the square root of the likelihood ratio is assigned the sign of the difference of the parameter estimate from its target:

$$r(\theta_0) = \text{sign}[\hat{\theta} - \theta_0]\sqrt{T} \quad (42)$$

to yield an asymptotically standard normal distribution of $r(\cdot)$. The theory of the signed root tests is given by Barndorff-Nielsen & Cox (1994), and examples of the practical use in social science applications were given by Andrews (2007).

By analogy, let us define the signed root of the scaled χ^2 difference test:

$$r_{\text{sc}}(\theta_0) = \text{sign}[\hat{\theta} - \theta_0]\sqrt{\bar{T}_d}. \quad (43)$$

To the extent that the distribution of \bar{T}_d can be approximated by χ_1^2 , the distribution of $r_{\text{sc}}(\cdot)$ can be approximated by the standard normal distribution. Hence, the one-sided signed root tests of (14) reject H_0 when, similarly to (16),

$$r(\theta_0) < z_\alpha \quad \text{or} \quad r_{\text{sc}}(\theta_0) < z_\alpha \quad (44)$$

Either the Bollen-Stine bootstrap or the Satorra-Bentler corrections enable researchers to test $H_0 : \theta_k = 0$ vs. $H_1 : \theta_k \neq 0$ even when the usual distributional assumptions are violated. However, both of these like the standard LR test assume that there is no structural misspecification. This is a limitation of these tests since if $\theta_k < 0$, then the model is structurally misspecified and the accuracy of any of the tests presented

in this subsection are open to question. To characterize the performance of this test, we need to be able to describe its distribution under the null (i.e., when $\theta_k = 0$) and under the alternative (i.e., when $\theta_k < 0$). Either of them may be tricky if we admit a possibility that the model is misspecified.

Let us now relate the signed root tests (42) and (43) to the tests at the boundary (39) and (40). For either of them to reject the null, the parameter estimate $\hat{\theta}_{k1}$ must be negative, and the (straight or scaled) likelihood ratio difference must be greater than the critical value $\chi_{1,0.95}^2$ for a 5% level test, say. Hence these tests are algebraically equivalent to one another, and application of either one will be dictated by the convenience of the estimation results manipulation by the researcher. Currently, no SEM software provides any of these tests, but they can be easily incorporated in general statistical software packages, such as Stata or R, that allow arbitrary transformations of estimation results along with the functions to compute the tail probabilities and quantiles of the basic statistical distributions.

Note also that since the Wald tests and likelihood ratio tests are asymptotically equivalent for the two-sided hypotheses (Buse 1982), their one-sided equivalence follows once the sign of the estimate is accounted for. Thus not only the signed roots and the tests at the boundary are identical in finite samples, but they are also close to the one-sided Wald test (16) in large samples.

3.5 Multiple and conditional tests of negative variances

In practice, a researcher will scan the output to detect inadmissible estimates and will only test those variances that are negative. A reviewer noted that this conditional procedure with many parameters in question may produce inaccurate rejection probabilities and p -values. First, potentially there are as many tests of Heywood cases, even though informal ones, as there are variance parameters in the model. Second, the formal tests, like the ones suggested above, are only performed when a negative estimate is observed, i.e., after the researcher had “peeked” at the data. We shall analyze these two issues in turn.

In our experience, multiple negative variances are possible, but rare in practice. On the face of it, a single negative error variance leads to a single significance tests and the issue of multiple testing does not emerge. However from another perspective an informal test for negative variances happens as a researcher scans the results of SEM to highlight only the negative values. One could argue that this scanning of values is an informal significance test in that positive variance values are determined to not be statistically significantly negative population variances. From this perspective, a researcher is performing as many significance tests as there are variances in the model and only performing formal tests when the variance estimate is negative. If we accept this argument, then we have a multiple testing problem where there are a large number of significant tests performed albeit most are informal. This multiple testing leads the usual Type I error probabilities of finding a significant negative variance to be higher than estimated with the usual single test probability. If, for instance, there are 20 variances in a model and the estimates for all but one are nonnegative and if we test the statistical significance for the negative one, the probability of finding it statistically significant is higher than the nominal Type I error probability since we “tested” (or checked) the other 19 variances to make sure they were positive. An implication of this is that we should set the Type I error probability lower than the usual 0.05 level when testing the negative error variances to reflect the number of variances that are tested to be negative. A Bonferroni or Holm correction for multiple testing could prove helpful especially when using the single parameter tests of statistical significance.

An advantage of the single tests of variances is that some of these are available for one-sided tests ($H_0 : \theta_k \geq 0$ vs. $H_1 : \theta_k < 0$) and in the case of the Wald test, the Huber sandwich asymptotic standard errors that underlie them apply even with structural misspecification. If a two-sided test of $H_0 : \theta_k = 0$ vs. $H_1 : \theta_k \neq 0$ applies, then there is another option to test multiple hypotheses. Using the appropriate variance-covariance matrix of the parameter estimator, a Wald test can be constructed as special cases of the general linear hypothesis $H_0 : C\theta = c$. If the estimates of θ are asymptotically normal, the Wald test statistic (Buse 1982, Amemiya 1985, Davidson & MacKinnon 1993, Ferguson 1996) is formed as a quadratic form involving the parameters and restrictions of interest:

$$W = (C\hat{\theta} - c)^T (C\widehat{V}C^T)^{-1} (C\hat{\theta} - c) \quad (45)$$

that has an asymptotic χ^2 distribution with the degrees of freedom equal to the effective number of hypotheses tested (rank of C). Note that this result, unlike the likelihood ratio χ^2 , does not depend on normality of underlying data, provided the estimator $\widehat{V}[\hat{\theta}]$ is consistent. On the other hand, Wald tests are known to be sensitive to alternative non-linear specifications (Lafontaine & White 1986, Phillips & Park 1988) and hence to re-parameterization (Gonzalez & Griffin 2001). The advantage of this simultaneous test is that it will test whether all variances are significantly different from zero. A disadvantage of it is that given its null hypothesis, it does not discriminate between variances that are positive and significantly different from zero from those that are negative and significant where the latter are of greater interest.

A second issue is the possible impact of the conditional nature of the test for positive vs. negative variance in the case of a single variance being tested. Does this conditioning on the sign of the initial estimate affect the size of the test? Let us consider a one-sided test with the asymptotic standard normal distribution of the test statistic T and a critical value c , such as Wald test (16), $T = \hat{\theta}/\text{s.e.}[\hat{\theta}]$, or a signed root test (44), $T = \text{sign}[\hat{\theta}]\sqrt{\widehat{T}_d}$. The probability that we reject $H_0 : \theta \geq 0$ is the unconditional probability, $\text{Prob}[T < c]$. If a researcher only tests H_0 when the sample variance is negative, then the resulting rejection probability is $\text{Prob}[T < c \cap T < 0] = \text{Prob}[T < c|T < 0]\text{Prob}[T < 0]$. The question becomes, what is the relation between $\text{Prob}[T < c]$ and $\text{Prob}[T < c \cap T < 0]$? Let us rewrite the latter probability with a different conditioning as $\text{Prob}[T < c \cap T < 0] = \text{Prob}[T < 0|T < c]\text{Prob}[T < c]$. We can see that if the first conditional probability in the RHS is equal to 1, then the rejection probability upon conducting an informal screening is the same as the total rejection probability $\text{Prob}[T < c]$. Obviously, if the critical value $c < 0$, then $\text{Prob}[T < 0|T < c] = 1$. The conditional probability of testing for negative error variances only if the variance is negative is equivalent to the unconditional probability of testing whether T is less than the critical value. The latter can be set to the desired level by choosing c to be the corresponding percentile of the standard normal distribution, and for test levels less than 50%, c will be negative for the test of $H_0 : \theta \geq 0$. No adjustment is needed.

In sum, we need to take account of the multiple tests implicitly performed when we screen the variances to check whether they are negative or positive. But we do not need to adjust our individual test when we only perform the test for negative error variances. Of course, we are assuming that the test statistics have the appropriate distribution for the data and model examined. In particular, the Wald tests must utilize consistent standard errors.

3.6 Summary of the tests

Table 2 summarizes the tests we are studying, and lists our expectations regarding their performance. We describe the anticipated performance of the tests under correct and incorrect structural specification. The anticipations under correct structural specification are based on what is already known in the literature, or as projected for the one-sided tests based on the known characteristics of the two-sided counterparts. We also control for the situations where the normal theory inference is applicable (“Normal” columns of the table), and when it is violated (“Non-normal” columns, i.e., excess kurtosis or lack of asymptotic robustness). The latter situation rules out the tests based on the normal theory, i.e., uncorrected likelihood ratio tests and Wald tests/CIs based on information matrix.

We indicate which of the tests we expect to perform well in large samples ($\longrightarrow \alpha$ and $\rightarrow \alpha$ symbols in the table), which of the tests we expect to demonstrate their asymptotic properties with relatively smaller sample sizes ($\rightarrow \alpha$ symbol), and which tests are expected to have wrong size ($\nrightarrow \alpha$ symbol). If a test has wrong size, it disqualifies it from consideration under the alternative (N/A symbol). If a test does achieve the target size, at least asymptotically, it is of interest to consider its power, so as to compare different types of tests (“power?” symbol).

Study of the confidence intervals allows for simultaneous assessment of levels of tests and their powers when the structure of a model is misspecified. The level of the test can be assessed by computing the fraction of times the confidence interval covers the true (negative) value, while the power of the test can be assessed by computing the fraction of times the confidence interval does not cover the null value of zero.

All versions of Wald tests are expected to have comparable power, provided that the standard errors are consistent under misspecification. We expect this consistency to hold for empirical sandwich and empirical bootstrap standard errors, and we expect the information matrix based, WLS and Bollen-Stine bootstrap standard errors to be biased down under misspecification. Two-sided tests are expected to have power lower than their one-sided counterparts. However no other comparisons of power can be made at this stage.

The large sample power of the local Wald test (16) can be found using asymptotic normality of the estimates. As

$$\sqrt{N}(\hat{\theta}_k - \theta_{0k}) \approx N(0, V_k),$$

where V_k is the asymptotic variance found as the appropriate diagonal element of the asymptotic variance-covariance matrix (24), we can find

$$\begin{aligned} & \text{Prob}[\hat{\theta}_k/\text{s.e.}[\hat{\theta}_k] < z_\alpha] \\ &= \text{Prob}[(\hat{\theta}_k - \theta_{0k})/\sqrt{V_k/N} \cdot \sqrt{V_k/N}/\text{s.e.}[\hat{\theta}_k] + \theta_{0k}/\sqrt{V_k/N} \cdot \sqrt{V_k/N}/\text{s.e.}[\hat{\theta}_k] < z_\alpha] \\ &\approx \text{Prob}[Z < z_\alpha - \theta_{0k}/\sqrt{V_k/N}] \equiv \Phi(z_\alpha - \theta_{0k}/\sqrt{V_k/N}) \end{aligned} \quad (46)$$

which increases to 1 as $N \rightarrow \infty$ since θ_{0k} is negative.

The one-sided LR-type tests, the test correcting for the boundary and the signed root test, are asymptotically equivalent to the Wald test. The same power calculation can be used for large samples.

Table 2: Tests of Heywood cases.

Structural specification: Distributional specification:	Correct		Misspecified	
	Normal	Non-normal	Normal	Non-normal
Wald tests				
Info matrix s.e. (17)	→ α	α	N/A	N/A
WLS s.e. (21)	→ α	→ α	N/A	N/A
Empirical sandwich s.e. (24)	→ α	→ α	power?	power?
Empirical bootstrap s.e. (25)	→ $\alpha(B)$	→ $\alpha(B)$	power?	power?
Bollen–Stine bootstrap s.e. (26)	→ $\alpha(B)$	→ $\alpha(B)$	N/A	N/A
Confidence intervals				
Info matrix s.e. (17)	→ α	α	α	α
WLS s.e. (21)	→ α	→ α	α	α
Empirical sandwich s.e. (24)	→ α	→ α	→ α power?	→ α power?
Empirical bootstrap s.e. (25)	→ $\alpha(B)$	→ $\alpha(B)$	→ $\alpha(B)$ power?	→ $\alpha(B)$ power?
Bollen–Stine bootstrap s.e. (26)	→ $\alpha(B)$	→ $\alpha(B)$	α	α
Empirical bootstrap percentile (30)	→ $\alpha(B)$	→ $\alpha(B)$	→ $\alpha(B)$ power?	→ $\alpha(B)$ power?
Bollen–Stine bootstrap percentile (31)	→ $\alpha(B)$	→ $\alpha(B)$	α	α
Likelihood ratio type tests				
<i>One-sided tests</i>				
Test at the boundary (39)	→ α	α	power?	N/A
Scaled test at the boundary (40)	→ α	→ α	power?	power?
Signed root of χ^2 -difference (42)	→ α	α	power?	N/A
Signed root of scaled difference (43)	→ α	→ α	power?	power?
<i>Two-sided tests</i>				
χ^2 -difference (41)	→ α	α	power?	N/A
Scaled difference (38)	→ α	→ α	power?	power?

Legend:

→ α : the test achieves the target level with moderate sample sizes

→ α : the test needs large sample sizes to achieve the target level

~~α~~ : the test violates assumptions needed for correct Type I level

(B): performance of the bootstrap scheme depends on the number of the bootstrap samples B

N/A: power of the test cannot be assessed as its size cannot be controlled for

4 Simulation study

Negative error variance in the population means that a model is structurally misspecified. However, negative error variance also can occur due to sampling fluctuations when the true variance is close to zero, and sample size is not very large. As Table 2 shows, there are numerous possible tests of negative error variances. Which of these tests work best in detecting that an error variance is significantly different from zero? To gauge the performance of different tests, we conducted simulations with the populations having Heywood cases in the misspecified model. For this to happen, we fit structurally misspecified models to realistic covariance matrices. In practical situations, researchers never know whether their models are true, and thus the possibility of structural misspecification is non-negligible. Other studies of negative variances have relied on simulations where the negative variances are due to sampling fluctuations. Our simulations are unique in that we design a study where a structural misspecification of a model creates a negative error variance in the population.

The simulation design includes two models: a saturated model with 3 observed variables, and an overidentified model with 4 observed variables, with modifications to allow for the null behavior with correctly specified models, and the alternative behavior with incorrectly specified models. We study several distributions (normal, heavy-tailed $t(5)$, and a non-elliptic distribution), a range of sample sizes (from 50 to 5000), and a variety of tests outlined in the previous section.

We are interested in several inference outcomes, both in terms of the behavior under the null (keeping the size of the test and coverage of the confidence intervals under control), and under the alternative (relative powers of the tests that have the correct size). A large number of replications per each combination of settings is required to estimate (small) probabilities of type I error. The use of the same data sets for estimation with different types of standard errors can be viewed as a variance reduction technique (Skrondal 2000).

All variables were generated with means of zeroes, and the models were estimated in deviation scores, so that the means and intercepts could be ignored. The simulations were performed in Stata SE statistical software (Stata Corp. 2007) using the estimation package `confa` for confirmatory factor analysis model estimation (Kolenikov 2009). Stata's internal maximum likelihood estimation routine `m1` yields the point estimates and the estimated covariance of the parameter estimates using one of the observed information, outer product of the gradients, and empirical sandwich estimator (referred in Stata as "robust" estimator; not to be confused with WLS standard errors (21), which are implemented in `confa` package with `vce(satorrabentler)` option). The necessary gradients are computed by Stata numerically, with a lot of attention given to computational accuracy (Gould, Pittblado & Sribney 2006).

A complete summary of the simulation results would involve $4 \text{ models} \times 3 \text{ distributions} \times 7 \text{ sample sizes} \times \left[8 \text{ parameters} \times (7 \text{ standard error estimates} + 7 \text{ associated Wald tests} + \text{characterization of bias} + 8 \text{ possible bootstrap tests}) + 4 \text{ likelihood ratio and related tests} \right] = 15,792$ characteristics to look at. An even greater number would arise if different measures of performance (e.g., mean, median, variance, quintiles, etc.) are used for these characteristics. Hence we are forced to present only a brief overview of the simulation results.

4.1 Three variables, saturated model

This example is designed to illustrate how using a false model might generate a Heywood case that is not present in the true model. The true data generating process is that of simultaneous equations:

$$\begin{aligned} \mathbb{V}[y_1] &= 1 \\ y_2 &= 0.3y_1 + \zeta_2, & \mathbb{V}[\zeta_2] &= 1, \\ y_3 &= 0.55y_1 + 0.8y_2 + \zeta_3, & \mathbb{V}[\zeta_3] &= 1, \end{aligned} \quad \Sigma = \begin{pmatrix} 1 & 0.3 & 0.79 \\ 0.3 & 1.09 & 1.037 \\ 0.79 & 1.037 & 2.264 \end{pmatrix} \quad (47)$$

Suppose a researcher incorrectly believes that a confirmatory factor analysis (CFA) model

$$y_k = \lambda_k \xi + \delta_k, \quad k = 1, 2, 3 \quad (48)$$

should be fitted to the data. To identify the scale of the latent variable, λ_1 is set to 1. The model is exactly identified, and using the population covariance matrix (47), the values of the parameters can be computed analytically to yield the population values

$$\lambda_2 = 1.313, \quad \lambda_3 = 3.457, \quad \phi_{11} = 0.229, \quad \theta_1 = 0.771, \quad \theta_2 = 0.696, \quad \theta_3 = -0.467$$

The Heywood case is observed with the third variable that has measurement error “variance” of -0.467 . FIML estimates will converge to those values in large samples⁷, and hence the probability of observing a Heywood case when CFA is fit to the data coming from (47) will approach 100% as $N \rightarrow \infty$.

The models are depicted on Fig. 1 using the path analysis conventions (Bollen 1989). The regression coefficients are shown on the arrows, and the variances of the exogenous variables, factors and errors are shown in angular brackets. Panel (a) shows the true model generating process. Panel (b) shows the population parameters of the structurally misspecified CFA model. Panel (c) shows the parameters of the model fitted with the variance of the last variable set to 0.

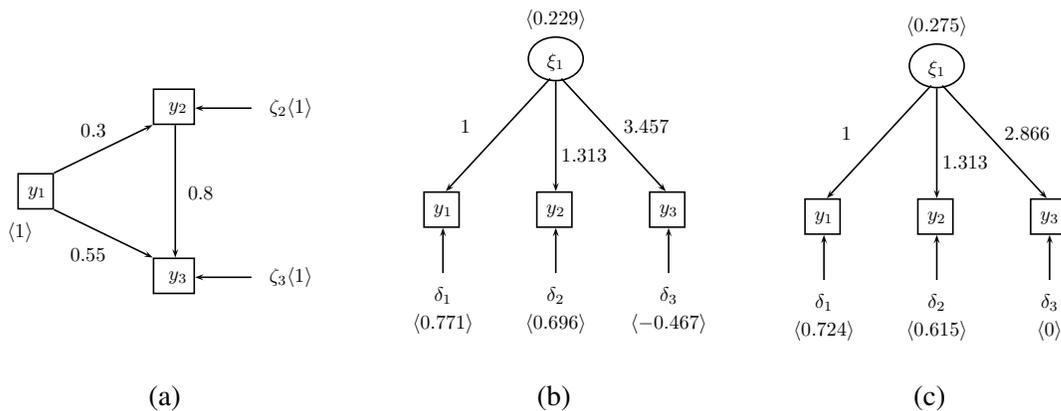


Figure 1: Population covariance structure with three variables: (a) data generating process, (b) fitted CFA model; (c) fitted CFA model with restricted variance. Population unique variances are in \langle angular brackets \rangle .

Table 3: Proportion of converged cases, Heywood case rate and relative bias of the “variance” parameter estimates for three variable CFA with multivariate normal and multivariate $t(5)$ distributions.

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Multivariate normal							
% Convergence	90.2%	96.8%	99.6%	99.9%	99.8%	98.3%	91.4%
% Heywood cases	71.3%	81.3%	90.6%	97.9%	99.8%	100%	100%
Relative bias, %							
$\hat{\theta}_1$	<i>-3.9</i>	<i>-1.8</i>	<i>-1.3</i>	-0.0	-0.2	-0.0	-0.1
$\hat{\theta}_2$	<i>-4.0</i>	<i>-1.8</i>	<i>-1.2</i>	-0.0	-0.2	-0.2	-0.0
$\hat{\theta}_3$	<i>-95.4</i>	<i>-51.2</i>	<i>-19.9</i>	<i>-9.0</i>	<i>-2.8</i>	-1.4	-0.3
Multivariate $t(5)$							
% Convergence	80.3%	90.6%	96.0%	98.8%	99.7%	97.8%	90.2%
% Heywood cases	65.7%	72.3%	81.2%	92.1%	97.6%	99.3%	99.8%
Relative bias, %							
$\hat{\theta}_1$	<i>-7.3</i>	<i>-3.9</i>	<i>-1.6</i>	<i>-0.3</i>	-0.4	-0.4	-0.2
$\hat{\theta}_2$	<i>-8.4</i>	<i>-3.9</i>	-0.7	0.2	-0.6	-0.2	0.0
$\hat{\theta}_3$	<i>-165</i>	<i>-94.8</i>	<i>-58.7</i>	<i>-25.2</i>	<i>-6.2</i>	<i>-5.1</i>	<i>-2.2</i>

Emphasis in italics indicates bias statistically significant at 5% level.

Sample sizes from 50 to 5000 were explored, with $R = 2000$ replications per setting. For each sample, we estimated the model with and without the constraint $\nabla[\delta_3] = 0$ using the population values of the parameters in Fig. 1(b) as the starting values to speed up convergence. Table 3 gives the convergence rate, Heywood case rate, and the bias of the error variance estimator when using the FIML estimator on the unrestricted CFA model in Figure 1(b). The results include variables generated from a multivariate normal distribution and a multivariate Student distribution with 5 degrees of freedom. Emphasis in italics shows statistically significant biases of parameter estimates. Some of these are substantively small (single digit percentages), but others are quite large.

Consider the samples from multivariate normal distributions first. The percentage of Monte Carlo samples where convergence was achieved ranges from 90% to 99.9%. As expected, the convergence rate is lowest at the smallest sample size ($N = 50$) and generally highest at the larger sample sizes.⁸ The percentage of Heywood cases (negative variances) increases with sample size with the lowest percent (71%) at $N = 50$ and essentially at 100% starting at $N = 1000$. We expect this since the population CFA model has a negative error variance for the y_3 variable due to the use of the structurally misspecified model.

⁸ The lower fractions of converged samples at $N \geq 2000$ can be explained by insufficient number of Newton-Raphson iterations specified in the defaults of the maximization procedure. This is not a serious limitation of the study; in the next section where a different model was studied, the limit was increased, and the issue was resolved.

We calculate the relative bias in Table 3 as

$$\text{Rel. bias}[\hat{\theta}_k] = \frac{\frac{1}{R} \sum_{i=1}^R \hat{\theta}_k^{(r)} - \theta_k}{\theta_k} \times 100\% \quad (49)$$

where $\hat{\theta}_k^{(r)}$ is the estimate of θ_k obtained in r -th simulated data set, and R is the number of Monte Carlo replications. The (downward) biases are less than 5% for the error variances of y_1 and y_2 which are positive in the population model in Figure 1 b. However, the percentage bias for the error variance in y_3 is very large at the smaller sample sizes (-95.4% to -19.9% for $N = 50$ to $N = 200$). The significance of bias was tested by using a z -test of

$$H_0 : \mathbb{E}[\hat{\theta}_{k,N}] = \theta_k \quad (50)$$

where the dependence of the distribution of $\hat{\theta}_k$ on the sample size N is made explicit with a subindex. This is a z -test where the Monte Carlo results are used as the data, and the RHS value for the parameter of interest is taken from Fig. 1(b).

Variables from distributions with excess multivariate kurtosis can lead to inaccurate normal-theory based standard errors (Browne 1984, Satorra 1992). To explore this issue we used the same model, but generated variables from a multivariate $t(5)$ distribution so as to create excess kurtosis leading to non-robust inference. The other parameters in the model were kept the same. The multivariate t distribution was obtained by sampling from the multivariate normal distribution first and dividing the resulting random data by $\sqrt{V_5/3}$, $V_5 \sim \chi_5^2$ independent of the data. Note that this “difficult” distribution is unfavorable for WLS standard errors. An estimate of the fourth order moments of the data is needed in (21), but it is not guaranteed to be consistent for this distribution. Similar patterns of results are found with these non-normal data, but the numeric values tend to indicate more severe problems. For instance, the proportion of converged cases is lower in the smaller sample sizes, and clear evidence of Heywood cases takes larger sample sizes to show. In addition, the relative bias percent is larger and for some parameters does not disappear even at $N = 5000$, although it keeps decreasing in the absolute value. The explanation is in the excess kurtosis of this distribution which negatively affects both the bias and the variance of the estimates in finite samples, even though the estimates are consistent in accordance with Browne (1984).

Verification of the conjectures made in Table 1 is performed in Table 4 for analytic standard errors and selected variance parameters. The upper half of Table 4 corresponds to the upper right corner of Table 1, and describes performance of various types of standard errors in misspecified models with normal data. Reported measures include the Monte Carlo means and medians of the standard errors, as well as the MSE of the standard errors as a stability measure (Kovar, Rao & Wu 1988):

$$\text{stability}[\hat{v}_t(\hat{\theta})] = \frac{\sum_{r=1}^R (\hat{v}_{t,r}^{1/2}[\hat{\theta}] - \mathbb{V}[\hat{\theta}]^{1/2})^2}{\mathbb{V}[\hat{\theta}]} \quad (51)$$

Here t indexes the different variance estimator types, $v_{t,r}$ is the estimated variance reported in r -th Monte Carlo replication, and $\mathbb{V}[\hat{\theta}] = 1/R \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2$ is the Monte Carlo variance of the parameter estimate $\hat{\theta}$. All types of standard errors are consistent, while information matrix standard errors are more stable. This is not surprising, as the IM standard errors are the maximum likelihood estimators, and hence are asymptotically efficient.

The bottom half of Table 4 corresponds to the lower right corner of Table 1, and describes performance of various types of standard errors in misspecified models with non-normal data. In this case, the information matrix based standard errors are biased down to the point of being useless. Besides, they become less stable as the sample size increases beyond $N = 500$ or 1000 , as the bias component dominates in the overall MSE in (51). The WLS and empirical sandwich standard errors exhibit great instability and skewness with low sample sizes, but approach the target standard deviation as the sample size increases.

Seemingly contrary to the prediction of Table 1, WLS standard errors performed reasonably well in both normal and non-normal situations. This happened because the model was saturated, and hence σ_* in (20) was identically zero. In this situation, WLS estimator (21) is appropriate.

The information matrix also performed well in the misspecified case with multivariate normal data. To explain this, we need to consider equation (20). There, the (asymptotic) normal theory weight matrix (11) is given by $V_N = .5D^-(\Sigma \otimes \Sigma)D^{-T}$, cancels with the variance of the moments $\Omega = 2D'(\Sigma \otimes \Sigma)D$, and after some trivial algebra, the asymptotic variance becomes the inverse of the information matrix, $1/N(\Delta V \Delta')^{-1}$.

Let us now turn to tests of Heywood cases described in Table 2. Only calculations of the analytical tests and standard errors were conducted for this model. Computationally intensive bootstrap simulations for both the empirical and Bollen–Stine bootstraps are conducted for another example and reported in the next section. Calculation of the Satorra-Bentler scaled differences and their boundary-aware versions failed for numerical reasons. When a saturated model is estimated without restrictions on θ_3 , computation of the scaling correction in (36) involves taking differences of two identical matrices resulting in a computer zero, and computation of the scaled test statistic involves division $0/0$. The results were numerically unstable.

The results for the remaining tests are reported in Tables 5–6. To study sizes of the tests, the data was simulated from the population shown on Fig. 1(c) with somewhat sparser grid of the sample sizes. If the coverage of a CI is indeed 95%, then with 2000 Monte Carlo replications, the number of intervals that fail to cover the true parameter value is a binomial random variable with $n = 2000$ and $p = 0.05$, and the Monte Carlo standard deviation of coverage is $\sqrt{0.95 \cdot 0.05/2000} = 0.49\%$. We consider performance within 2% of the target 5% to be adequate. Such a high value is necessary to correct for multiple testing: each simulation results table we present reports about 100 test results. Performance outside of the 2% accuracy margin is indicated with *italics*. These italicized results raise questions about accuracy of the test for the sample size, data distribution and model specification.

Expectations put forward in Table 2 are by and large confirmed. All the tests achieve their target sizes in large samples. Poor performance of Wald tests and confidence intervals in small samples is likely due to biases of parameter estimates, as the sampling distributions of $\hat{\theta}_3$ have wrong centering. All other tests demonstrate accurate performance. In particular, all likelihood ratio type tests have the right size in samples as small as $N = 100$, which exceeded our expectations regarding these tests. However, this performance critically hinges on the assumption of multivariate normality (or, more generally speaking, asymptotic robustness). Once it is violated, as is the case with $t(5)$ data, the likelihood ratio type tests break down. Likewise, the Wald tests and confidence intervals based on the normal theory information matrix standard errors are clearly inappropriate. Wald tests with either WLS or empirical sandwich standard errors achieve their target sizes asymptotically, although the sample sizes required for that exceed $N = 1000$. Confidence intervals testing method shows good performance with some parameters, but requires large sizes for oth-

Table 4: Standards errors in the three variable confirmatory factor analysis model of Figure 1(b) (incorrectly specified model).

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Multivariate normal data							
Heywood case: $\hat{\theta}_3$ (population value $\theta_3 = -0.4667$)							
MC s.d.	0.9477	0.7773	0.5221	0.3004	0.2057	0.1381	0.0849
IM s.e.: Mean	1.0698	0.7680	0.5014	0.2952	0.2018	0.1401	0.0878
Median	0.7532	0.5814	0.4287	0.2755	0.1950	0.1385	0.0873
Stability	1.4002	0.6917	0.2946	0.0892	0.0388	0.0181	0.0071
WLS s.e.: Mean	1.1999	0.8081	0.5035	0.2951	0.2015	0.1398	0.0877
Median	0.7578	0.5814	0.4259	0.2754	0.1942	0.1380	0.0872
Stability	2.1096	0.8502	0.3403	0.0922	0.0403	0.0185	0.0077
ES s.e.: Mean	1.2075	0.8104	0.5040	0.2951	0.2016	0.1399	0.0877
Median	0.7581	0.5814	0.4259	0.2754	0.1942	0.1380	0.0872
Stability	2.2317	0.8747	0.3477	0.0925	0.0404	0.0185	0.0082
Non-Heywood case: $\hat{\theta}_1$ (population value $\theta_1 = 0.7715$)							
MC s.d.	0.1634	0.1190	0.0839	0.0530	0.0379	0.0269	0.0166
IM s.e.: Mean	0.1648	0.1190	0.0846	0.0540	0.0381	0.0270	0.0170
Median	0.1624	0.1179	0.0841	0.0539	0.0381	0.0270	0.0170
Stability	0.0474	0.0232	0.0118	0.0050	0.0024	0.0012	0.0014
WLS s.e.: Mean	0.1607	0.1171	0.0838	0.0539	0.0381	0.0269	0.0170
Median	0.1554	0.1152	0.0831	0.0537	0.0380	0.0269	0.0170
Stability	0.0665	0.0349	0.0184	0.0079	0.0039	0.0019	0.0017
ES s.e.: Mean	0.1607	0.1172	0.0838	0.0539	0.0381	0.0269	0.0170
Median	0.1554	0.1152	0.0831	0.0537	0.0380	0.0269	0.0170
Stability	0.0667	0.0350	0.0185	0.0079	0.0039	0.0019	0.0017
Multivariate $t(5)$ data							
Heywood case: $\hat{\theta}_3$ (population value $\theta_3 = -0.4667$)							
MC s.d.	1.1327	0.9801	0.8127	0.5399	0.3039	0.2393	0.1530
IM s.e.: Mean	0.9769	0.7321	0.5403	0.3189	0.2068	0.1433	0.0886
Median	0.6278	0.5224	0.4199	0.2724	0.1926	0.1374	0.0870
Stability	0.8692	0.4871	0.3587	0.2660	0.1479	0.1808	0.1865
ES s.e.: Mean	1.4558	1.0963	0.8249	0.4873	0.3086	0.2257	0.1443
Median	0.7333	0.6527	0.5456	0.3821	0.2761	0.2026	0.1336
Stability	3.7408	2.1162	1.5692	0.7391	0.1847	0.2604	0.1149
WLS s.e.: Mean	1.4417	1.0903	0.8200	0.4866	0.3085	0.2256	0.1443
Median	0.7333	0.6526	0.5455	0.3820	0.2761	0.2026	0.1336
Stability	3.4798	2.0353	1.4697	0.7223	0.1842	0.2595	0.1147
Non-Heywood case: $\hat{\theta}_1$ (population value $\theta_1 = 0.7715$)							
MC s.d.	0.2243	0.1788	0.1336	0.0939	0.0640	0.0466	0.0320
IM s.e.: Mean	0.1556	0.1153	0.0841	0.0539	0.0381	0.0269	0.0170
Median	0.1479	0.1114	0.0822	0.0531	0.0379	0.0268	0.0170
Stability	0.1366	0.1447	0.1471	0.1778	0.1684	0.1804	0.2132
ES s.e.: Mean	0.1959	0.1610	0.1249	0.0863	0.0626	0.0458	0.0304
Median	0.1685	0.1406	0.1113	0.0780	0.0581	0.0425	0.0281
Stability	0.2986	0.2786	0.1673	0.1383	0.0971	0.0932	0.0849
WLS s.e.: Mean	0.1957	0.1610	0.1248	0.0863	0.0626	0.0458	0.0304
Median	0.1685	0.1406	0.1113	0.0779	0.0581	0.0425	0.0281
Stability	0.2956	0.2781	0.1668	0.1381	0.0971	0.0932	0.0848

Table 5: Test sizes in the three variable confirmatory factor analysis of Figure 1(c) (correctly specified model).

	Sample size			
	100	300	1000	3000
Multivariate normal data				
One-sided 5% Wald tests				
Info matrix s.e. (17)	<i>0.000</i>	<i>1.287</i>	3.311	4.494
WLS s.e. (21)	<i>0.000</i>	<i>1.287</i>	3.311	4.390
Empirical sandwich s.e. (24)	<i>0.000</i>	<i>1.287</i>	3.311	4.390
Two-sided 95% confidence intervals: θ_3, Heywood case under alternative				
Info matrix s.e. (17)	94.166	95.545	95.648	95.017
WLS s.e. (21)	94.176	95.743	95.364	95.171
Empirical sandwich s.e. (24)	94.176	95.743	95.364	95.220
Two-sided 95% confidence intervals: θ_1, no Heywood case				
Info matrix s.e. (17)	<i>92.182</i>	94.604	94.560	95.017
WLS s.e. (21)	<i>91.147</i>	94.356	94.371	94.829
Empirical sandwich s.e. (24)	<i>91.147</i>	94.356	94.371	94.829
Likelihood ratio type tests				
<i>One-sided 5% tests</i>				
Test at the boundary (39)	4.434	5.099	4.967	5.227
Signed root of χ^2 -difference (42)	4.434	5.099	4.967	5.227
<i>Two-sided 5% tests</i>				
χ^2 -difference (41)	5.018	5.347	4.967	5.129
Multivariate $t(5)$ data				
One-sided 5% Wald tests				
Info matrix s.e. (17)	<i>0.382</i>	5.226	<i>11.630</i>	<i>12.023</i>
WLS s.e. (21)	<i>0.282</i>	<i>0.605</i>	<i>2.482</i>	3.375
Empirical sandwich s.e. (24)	<i>0.282</i>	<i>0.605</i>	<i>2.433</i>	3.375
Two-sided 95% confidence intervals: θ_3, Heywood case under alternative				
Info matrix s.e. (17)	<i>87.052</i>	<i>86.094</i>	<i>81.800</i>	<i>80.186</i>
WLS s.e. (21)	93.133	94.832	95.426	95.379
Empirical sandwich s.e. (24)	93.133	94.832	95.426	95.379
Two-sided 95% confidence intervals: θ_1, no Heywood case				
Info matrix s.e. (17)	<i>76.254</i>	<i>75.642</i>	<i>72.847</i>	<i>72.755</i>
WLS s.e. (21)	<i>83.490</i>	<i>88.780</i>	<i>92.019</i>	93.406
Empirical sandwich s.e. (24)	<i>83.490</i>	<i>88.780</i>	<i>92.019</i>	93.406
Likelihood ratio type tests				
<i>One-sided 5% tests</i>				
Test at the boundary (39)	<i>10.272</i>	<i>12.273</i>	<i>14.550</i>	<i>13.777</i>
Signed root of χ^2 -difference (42)	<i>10.272</i>	<i>12.273</i>	<i>14.550</i>	<i>13.777</i>
<i>Two-sided 5% tests</i>				
χ^2 -difference (41)	<i>15.719</i>	<i>18.245</i>	<i>19.854</i>	<i>20.330</i>

Emphasis in italics indicates test size different from 5% or coverage different from 95%.

ers. As a conservative suggestion, the asymptotic tests can be trusted with sample sizes $N > 1000$ in a non-normal situation for the current model size.

Table 6 compares the power of the tests. The combinations of sample sizes and distribution that were identified as problematic for the test size in Table 5 are repeatedly highlighted in Table 6. In the multivariate normal situation, the best tests are the ones based on one-sided likelihood ratio tests (the boundary modification and the signed root). They were shown to maintain accurate sizes for all sample sizes studied, and they also have the greatest power of all comparable tests. The power of the one-sided Wald tests is slightly smaller, and they need sample sizes above $N = 1000$ to have the right size. There are no discernible differences between Wald tests based on three different variance estimators (information matrix, WLS and empirical sandwich standard errors). Two-sided tests (based on likelihood ratio chi-square difference as well as confidence intervals) have lower power, with likelihood ratio performing a little better. With multivariate $t(5)$ data, only Wald tests and confidence intervals based on WLS or empirical sandwich standard errors showed appropriate test sizes in Table 5. Hence the normal theory inference results are excluded from consideration. The only meaningful power comparison is between the one-sided and two-sided tests. One-sided tests have an advantage for the sample sizes $N \geq 2000$ when such a comparison can be made.

Overall, when the model is saturated and the data are multivariate normal, one-sided modifications of the chi-square difference tests are the recommended tests. One-sided Wald tests or confidence intervals also have good power, but they only produce accurate test sizes in samples larger than $N \geq 500$. When the normality assumption is violated, the only means to construct Wald tests (or equivalent confidence intervals) are the WLS or empirical sandwich standard errors. One-sided intervals are preferred for one-sided testing situations, but they require sample sizes $N \geq 2000$ to ensure an accurate test size. Two-sided intervals have lower power, but they also work in smaller sample sizes ($N \geq 500$).

As a conclusion for this model, it should be noted that proper testing of the Heywood case allows us to establish that the model is inappropriate even when the overall fit test is not available due to the model being exactly identified. We are not aware of any other approaches to test model fit in those situations.

Table 6: Power of the tests in the three variable confirmatory factor analysis of Figure 1(b), (incorrectly specified model).

	Sample size						
	50	100	200	500	1000	2000	5000
Multivariate normal data							
One-sided Wald tests							
Info matrix s.e. (17)	<i>0.000</i>	<i>0.309</i>	<i>7.781</i>	53.677	85.159	99.238	100
WLS s.e. (21)	<i>0.107</i>	<i>0.406</i>	<i>8.763</i>	53.777	85.293	99.290	100
Empirical sandwich s.e. (24)	<i>0.107</i>	<i>0.355</i>	<i>8.763</i>	53.777	85.293	99.290	100
Two-sided confidence intervals: θ_3, Heywood case, power							
Info matrix s.e. (17)	<i>1.598</i>	<i>0.566</i>	1.155	31.066	72.809	97.815	100
WLS s.e. (21)	<i>1.974</i>	<i>0.660</i>	2.203	31.916	72.186	97.719	100
Empirical sandwich s.e. (24)	<i>1.974</i>	<i>0.660</i>	2.203	31.866	72.186	97.719	100
Two-sided confidence intervals: θ_3, Heywood case, coverage							
Info matrix s.e. (17)	<i>89.477</i>	<i>91.662</i>	94.026	95.348	94.970	95.681	94.809
WLS s.e. (21)	<i>88.794</i>	<i>91.066</i>	94.542	95.148	94.547	95.590	95.455
Empirical sandwich s.e. (24)	<i>88.794</i>	<i>91.066</i>	94.542	95.198	94.547	95.590	95.455
Confidence intervals: θ_1, no Heywood case, coverage							
Info matrix s.e. (17)	<i>89.752</i>	<i>92.795</i>	<i>92.671</i>	95.398	94.920	94.461	95.628
WLS s.e. (21)	<i>88.794</i>	<i>92.284</i>	<i>92.489</i>	95.098	94.847	94.425	95.507
Empirical sandwich s.e. (24)	<i>88.794</i>	<i>91.066</i>	<i>94.542</i>	95.198	94.547	95.590	95.455
Likelihood ratio type tests							
<i>One-sided tests</i>							
Test at the boundary (39)	10.193	22.542	37.851	68.034	89.243	99.644	100
Signed root of χ^2 -difference (42)	10.193	22.542	37.851	68.034	89.243	99.644	100
<i>Two-sided tests</i>							
χ^2 -difference (41)	5.234	14.720	25.954	56.028	82.221	98.577	100
Multivariate $t(5)$ data							
One-sided Wald tests							
WLS s.e. (21)	<i>0.402</i>	<i>0.585</i>	<i>2.558</i>	<i>18.882</i>	<i>48.692</i>	79.241	97.446
Empirical sandwich s.e. (24)	<i>0.402</i>	<i>0.585</i>	<i>2.558</i>	<i>18.882</i>	<i>48.692</i>	79.241	97.446
Two-sided confidence intervals: θ_3, Heywood case, power							
WLS s.e. (21)	<i>3.042</i>	<i>1.915</i>	<i>0.972</i>	5.891	27.515	65.197	94.725
Empirical sandwich s.e. (24)	<i>3.042</i>	<i>1.915</i>	<i>0.972</i>	5.891	27.515	65.197	94.725
Two-sided confidence intervals: θ_3, Heywood case, coverage							
WLS s.e. (21)	<i>84.615</i>	<i>88.085</i>	<i>90.435</i>	93.303	94.970	94.772	94.392
Empirical sandwich s.e. (24)	<i>84.615</i>	<i>88.085</i>	<i>90.435</i>	93.303	94.970	94.772	94.392
Confidence intervals: θ_1, no Heywood case, coverage							
WLS s.e. (21)	<i>84.615</i>	<i>88.085</i>	<i>90.435</i>	93.303	94.970	94.772	94.392
Empirical sandwich s.e. (24)	<i>84.615</i>	<i>88.085</i>	<i>90.435</i>	93.303	94.970	94.772	94.392

Emphasis in italics indicates test size different from 5% or coverage different from 95%.

4.2 Four variables, overidentified model

To study the performance of the tests from Table 2 in an overidentified model, we used the population represented in Fig. 2. Like in the population of Fig. 1, the true data generating model is that of recursive regressions:

$$\begin{aligned}
 y_2 &= 0.9911y_1 + \zeta_2, & \mathbb{V}[y_1] &= 1, \\
 y_3 &= 1.05y_1 - 0.15y_2 + \zeta_3, & \mathbb{V}[\zeta_2] &= 1, \\
 y_4 &= 1.3y_1 + 0.1y_2 + \zeta_4, & \mathbb{V}[\zeta_3] &= 0.8, \\
 & & \mathbb{V}[\zeta_4] &= 0.7,
 \end{aligned}
 \quad \Sigma = \begin{pmatrix} 1 & 0.9911 & 0.9013 & 1.3911 \\ 0.9911 & 1.9823 & 0.7433 & 1.4867 \\ 0.9013 & 0.7433 & 1.6349 & 1.2906 \\ 1.3991 & 1.4867 & 1.2461 & 2.6675 \end{pmatrix} \quad (52)$$

If the researcher studying the data fits an improperly specified confirmatory factor analysis model, the population value of the variance of δ_1 is exactly 0. This represents a quasi-null situation. The true model (panel (a)) is still that of recursive regressions (52), but the fitted CFA model does not produce an improper solution: the population value of ψ_1 is zero rather than negative. Hence this represents the null situation in terms of the Heywood case setup (14), but an alternative situation in terms of the overall test (6) of correctly vs. incorrectly specified structure of a model.

We generated the simulation data for this scenario using the multivariate normal and multivariate $t(5)$ distributions. An additional non-elliptic distribution was also studied, in which ζ_2 had log-normal distribution with parameters $\mu = 0.5043$ and $\sigma = 0.5$. This log-normal distribution has variance equal to 1, moderate skewness equal to 1.750 and sizeable kurtosis equal to 8.898. Through the data generating regression model, non-normality is propagated to the variables y_2 , y_3 and y_4 .

Table 7 gives the descriptive statistics of the simulation. There were no convergence problems with any of the distributions or sample sizes. The proportion of Heywood cases is about 50% across different sample sizes. All estimates show biases in small samples. Interestingly, when the data were generated according to the regression model with non-normal errors, the bias was not monotone, but rather was oscillating, with sample sizes $N = 100$ and $N = 2000$ exhibiting significantly positive bias, and other sample sizes below $N < 1000$ exhibiting significantly negative bias.

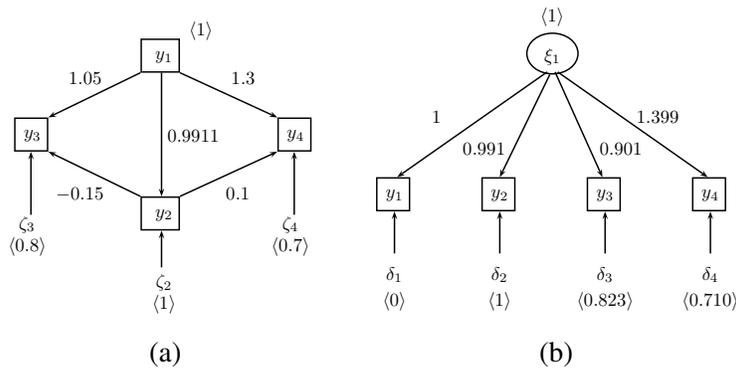


Figure 2: Quasi-null covariance structure with four variables: (a) data generating process, (b) fitted CFA model.

Table 7: Proportion of converged cases, Heywood case rate and bias of the “variance” parameter estimates for incorrectly specified four variable confirmatory factor analysis model of Fig. 2(b).

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Multivariate normal							
% Convergence	100%	100%	100%	100%	100%	100%	100%
% Heywood cases	49.9%	51.1%	51.1%	49.8%	49.3%	49.7%	50.2%
Bias							
$\hat{\theta}_1$	-0.0018	<i>-0.0023</i>	<i>-0.0018</i>	0.0001	0.0000	0.0001	-0.0001
$\hat{\theta}_2$	<i>-0.0445</i>	<i>-0.0204</i>	<i>-0.0088</i>	<i>-0.0039</i>	<i>-0.0027</i>	-0.0010	-0.0005
$\hat{\theta}_3$	<i>-0.0270</i>	<i>-0.0150</i>	<i>-0.0079</i>	<i>-0.0036</i>	-0.0012	<i>-0.0016</i>	0.0001
$\hat{\theta}_4$	<i>-0.0290</i>	<i>-0.0153</i>	-0.0052	<i>-0.0039</i>	-0.0002	-0.0007	0.0000
Multivariate $t(5)$							
% Convergence	100%	100%	100%	99.97%	100%	100%	100%
% Heywood cases	53.5%	50.5%	50.8%	50.3%	50.3%	50.3%	50.5%
Bias							
$\hat{\theta}_1$	<i>-0.0128</i>	<i>-0.0044</i>	-0.0018	-0.0006	-0.0003	-0.0004	-0.0001
$\hat{\theta}_2$	<i>-0.0607</i>	<i>-0.0296</i>	<i>-0.0116</i>	<i>-0.0080</i>	<i>-0.0050</i>	-0.0022	0.0001
$\hat{\theta}_3$	<i>-0.0377</i>	<i>-0.0293</i>	-0.0060	<i>-0.0064</i>	-0.0001	-0.0010	-0.0005
$\hat{\theta}_4$	<i>-0.0301</i>	<i>-0.0263</i>	<i>-0.0155</i>	<i>-0.0064</i>	-0.0025	-0.0004	-0.0001
Lognormal ζ_2 data							
% Convergence	99.96%	100%	100%	100%	100%	100%	100%
% Heywood cases	54.8%	47.9%	51.9%	51.6%	48.8%	44.2%	50.4%
Bias							
$\hat{\theta}_1$	<i>-0.0079</i>	<i>0.0022</i>	<i>-0.0017</i>	<i>-0.0014</i>	-0.0019	<i>0.0013</i>	-0.0002
$\hat{\theta}_2$	<i>-0.0298</i>	<i>-0.0240</i>	-0.0114	-0.0019	-0.0030	-0.0021	-0.0011
$\hat{\theta}_3$	<i>-0.0199</i>	<i>-0.0158</i>	<i>-0.0108</i>	<i>-0.0054</i>	-0.0010	<i>-0.0016</i>	-0.0001
$\hat{\theta}_4$	<i>-0.0154</i>	<i>-0.0172</i>	-0.0044	-0.0023	-0.0019	<i>-0.0034</i>	-0.0007

Emphasis in italics indicates bias statistically significant at 1% level.

4.2.1 Standard errors

Table 8 reports the summary of the standard errors for the normal distribution. All types of standard errors are biased down and asymmetric, especially with smaller samples. The bias is evident from comparisons of the mean reported standard error (the first line in each of the blocks) with the Monte Carlo standard deviation of the parameter estimates (top line of the table). Asymmetry of the distribution is evident from comparisons of the means and medians of the standard error distributions. The empirical sandwich standard errors are about 60% less stable than the information matrix based standard errors. Stability of the bootstrap standard errors depends on the number of the bootstrap samples, with greater number producing more stable results. The empirical bootstrap standard errors tend to be less stable than the empirical sandwich standard errors. To obtain comparable stability, the number of the bootstrap replications must notably exceed the sample size. While the Bollen-Stine bootstrap standard errors appear to perform well in smaller samples, they have a more pronounced downward bias in the larger samples. As a result their stability in the large samples is no better than that of the empirical bootstrap standard errors.

With non-normal elliptical Student $t(5)$ data (Table 9), the variances of the estimates are larger than in the normal case. However, the normal theory information matrix standard errors fail to account for that, demonstrating 30–50% downward asymptotic bias, and even greater bias in small samples. The WLS standard errors are asymptotically biased for the Heywood case, but not for the positive variance parameter. The factors behind this bias are explained in Section 3.1.2. The empirical bootstrap standard errors have lower bias and better stability than the empirical sandwich standard errors. The Bollen-Stine bootstrap standard errors are comparable to the WLS standard errors, also suffering from asymptotic bias. Hence, based on this Table, the best performing method to compute standard errors when the data do not satisfy the asymptotic robustness conditions appears to be the empirical bootstrap.

Table 10 shows the standard errors obtained when the data were generated with non-normal errors. The standard errors for this distribution closely resemble those for the normal distribution. Hence it should be concluded that the conditions of asymptotic robustness are satisfied for this distribution even though the model is structurally misspecified. The patterns of performance of different types of standard errors resembles the other two cases: the WLS and the Bollen-Stine bootstrap standard errors are biased down in larger samples, while the empirical bootstrap standard errors can offer performance that beats other variance estimation methods when the number of bootstrap samples is at least twice the sample size.

4.2.2 Test sizes

The next three tables describe the performance of the tests outlined in Table 2 in terms of their size. Using these tables, we want to identify the tests that have their target size of 5% under a wide variety of settings. Table 11 reports the results for the multivariate normal case. Except for a few tests that have problems at the smallest sample size ($N = 50$, scaled and adjusted overall tests and their one-sided versions), most tests appear to perform well. Table 12 reports the summary of the test sizes for multivariate $t(5)$ data. In this situation, all of the normal theory based inferential procedures (Wald tests and confidence intervals based on the information matrix standard errors; normal theory chi-square, chi-square difference and its one-sided versions) break down completely. The small sample problems of the scaled and adjusted chi-square tests and the scaled difference tests affect samples of the size up to $N \leq 200$. The bootstrap percentile tests

Table 8: Standards errors in four variable confirmatory factor analysis model of Figure 2(b), multivariate normal data, incorrectly specified model with zero population variance of δ_1 .

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Heywood case: $\hat{\theta}_1$ (population value $\theta_1 = 0$)							
MC s.d.	0.0636	0.0432	0.0302	0.0190	0.0135	0.0094	0.0061
IM s.e.:	Mean	0.0598	0.0419	0.0297	0.0188	0.0133	0.0094
	Median	0.0576	0.0412	0.0295	0.0188	0.0133	0.0094
	Stability	0.0387	0.0192	0.0098	0.0039	0.0020	0.0011
WLS s.e.:	Mean	0.0503	0.0365	0.0261	0.0166	0.0118	0.0084
	Median	0.0484	0.0359	0.0259	0.0166	0.0118	0.0084
	Stability	0.0583	0.0293	0.0161	0.0063	0.0039	0.0026
ES s.e.:	Mean	0.0608	0.0425	0.0300	0.0190	0.0134	0.0095
	Median	0.0577	0.0415	0.0296	0.0189	0.0134	0.0095
	Stability	0.0598	0.0306	0.0170	0.0065	0.0035	0.0018
EB s.e. ($B = 60$):	Mean	0.0659		0.0300		0.0133	
	Median	0.0608		0.0297		0.0133	
	Stability	0.0705		0.0246		0.0111	
EB s.e. ($B = 300$):	Mean	0.0659		0.0302		0.0134	
	Median	0.0616		0.0295		0.0133	
	Stability	0.0540		0.0191		0.0056	
BSBPF s.e. ($B = 60$):	Mean	0.0554		0.0275		0.0124	
	Median	0.0531		0.0273		0.0125	
	Stability	0.0671		0.0257		0.0134	
BSBPF s.e. ($B = 300$):	Mean	0.0576		0.0278		0.0124	
	Median	0.0539		0.0274		0.0124	
	Stability	0.0636		0.0183		0.0061	
BSBZV s.e. ($B = 60$):	Mean	0.0543		0.0272		0.0124	
	Median	0.0529		0.0270		0.0124	
	Stability	0.0656		0.0246		0.0132	
BSBZV s.e. ($B = 300$):	Mean	0.0537		0.0274		0.0124	
	Median	0.0517		0.0272		0.0124	
	Stability	0.0529		0.0163		0.0072	

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

Table 8: (continued).

Characteristic	Sample sizes							
	50	100	200	500	1000	2000	5000	
Non-Heywood case: $\hat{\theta}_3$ (population value $\theta_3 = 0.823$)								
MC s.d.	0.1671	0.1167	0.0847	0.0531	0.0370	0.0260	0.0165	
IM s.e.:	Mean	0.1633	0.1164	0.0828	0.0527	0.0373	0.0264	0.0167
	Median	0.1601	0.1158	0.0823	0.0526	0.0373	0.0264	0.0167
	Stability	0.1671	0.1167	0.0847	0.0531	0.0370	0.0260	0.0165
WLS s.e.:	Mean	0.1551	0.1146	0.0832	0.0532	0.0378	0.0268	0.0170
	Median	0.1516	0.1130	0.0826	0.0531	0.0377	0.0268	0.0170
	Stability	0.1671	0.1167	0.0847	0.0531	0.0370	0.0260	0.0165
ES s.e.:	Mean	0.1594	0.1150	0.0825	0.0525	0.0372	0.0264	0.0167
	Median	0.1551	0.1128	0.0816	0.0523	0.0371	0.0263	0.0167
	Stability	0.1671	0.1167	0.0847	0.0531	0.0370	0.0260	0.0165
EB s.e. ($B = 60$):	Mean	0.1618		0.0809		0.0372		
	Median	0.1551		0.0797		0.0372		
	Stability	0.1723		0.0861		0.0369		
EB s.e. ($B = 300$):	Mean	0.1591		0.0821		0.0371		
	Median	0.1561		0.0808		0.0371		
	Stability	0.1723		0.0862		0.0369		
BSBPF s.e. ($B = 60$):	Mean	0.1518		0.0824		0.0375		
	Median	0.1494		0.0819		0.0375		
	Stability	0.1723		0.0862		0.0369		
BSBPF s.e. ($B = 300$):	Mean	0.1569		0.0831		0.0379		
	Median	0.1537		0.0825		0.0377		
	Stability	0.1723		0.0862		0.0369		
BSBZV s.e. ($B = 60$):	Mean	0.1561		0.0834		0.0376		
	Median	0.1529		0.0834		0.0374		
	Stability	0.1655		0.0843		0.0368		
BSBZV s.e. ($B = 300$):	Mean	0.1567		0.0832		0.0381		
	Median	0.1541		0.0830		0.0378		
	Stability	0.1655		0.0843		0.0368		

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

Table 9: Standards errors in four variable confirmatory factor analysis model of Figure 2(b), Student $t(5)$ data, incorrectly specified model with zero population variance of δ_1 .

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Heywood case: $\hat{\theta}_1$ (population value $\theta_1 = 0$)							
MC s.d.	0.1098	0.0880	0.0475	0.0300	0.0225	0.0152	0.0101
IM s.e.:	Mean	0.0632	0.0435	0.0299	0.0188	0.0133	0.0094
	Median	0.0555	0.0402	0.0290	0.0185	0.0132	0.0094
	Stability	0.2100	0.2018	0.2146	0.2550	0.2289	0.2423
WLS s.e.:	Mean	0.0658	0.0463	0.0357	0.0243	0.0179	0.0129
	Median	0.0530	0.0417	0.0324	0.0225	0.0166	0.0122
	Stability	0.2398	0.1905	0.1805	0.1786	0.1302	0.1388
ES s.e.:	Mean	0.0851	0.0564	0.0418	0.0279	0.0205	0.0147
	Median	0.0655	0.0493	0.0375	0.0258	0.0189	0.0138
	Stability	0.3265	0.2524	0.2427	0.2569	0.1740	0.1669
EB s.e. ($B = 60$):	Mean	0.0917		0.0425		0.0201	
	Median	0.0705		0.0374		0.0189	
	Stability	0.2457		0.1765		0.1248	
EB s.e. ($B = 300$):	Mean	0.0997		0.0450		0.0206	
	Median	0.0738		0.0384		0.0193	
	Stability	0.2523		0.2120		0.0948	
BSBPF s.e. ($B = 60$):	Mean	0.0765		0.0375		0.0188	
	Median	0.0576		0.0336		0.0176	
	Stability	0.2498		0.1326		0.1242	
BSBPF s.e. ($B = 300$):	Mean	0.0764		0.0392		0.0193	
	Median	0.0594		0.0338		0.0178	
	Stability	0.2641		0.1830		0.1781	
BSBZV s.e. ($B = 60$):	Mean	0.0675		0.0375		0.0184	
	Median	0.0579		0.0341		0.0175	
	Stability	0.2974		0.1750		0.1634	
BSBZV s.e. ($B = 300$):	Mean	0.0645		0.0365		0.0188	
	Median	0.0554		0.0345		0.0177	
	Stability	0.2314		0.1591		0.1700	

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

Table 9: (continued).

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Non-Heywood case: $\hat{\theta}_3$ (population value $\theta_3 = 0.823$)							
MC s.d.	0.2743	0.2004	0.1522	0.1053	0.0714	0.0519	0.0317
IM s.e.:	Mean	0.1613	0.1149	0.0833	0.0525	0.0374	0.0264
	Median	0.1501	0.1111	0.0814	0.0519	0.0370	0.0263
	Stability	0.2743	0.2004	0.1522	0.1053	0.0714	0.0519
WLS s.e.:	Mean	0.1880	0.1501	0.1214	0.0834	0.0626	0.0457
	Median	0.1623	0.1344	0.1079	0.0758	0.0571	0.0421
	Stability	0.2743	0.1988	0.1522	0.1053	0.0714	0.0519
ES s.e.:	Mean	0.2153	0.1632	0.1285	0.0861	0.0640	0.0464
	Median	0.1779	0.1417	0.1117	0.0773	0.0578	0.0423
	Stability	0.2743	0.1988	0.1522	0.1053	0.0714	0.0519
EB s.e. ($B = 60$):	Mean	0.2121		0.1238		0.0622	
	Median	0.1800		0.1087		0.0573	
	Stability	0.2715		0.1499		0.0712	
EB s.e. ($B = 300$):	Mean	0.2082		0.1288		0.0624	
	Median	0.1775		0.1132		0.0577	
	Stability	0.2688		0.1499		0.0713	
BSBPF s.e. ($B = 60$):	Mean	0.1912		0.1173		0.0637	
	Median	0.1603		0.1065		0.0590	
	Stability	0.2682		0.1498		0.0713	
BSBPF s.e. ($B = 300$):	Mean	0.1999		0.1218		0.0657	
	Median	0.1729		0.1083		0.0583	
	Stability	0.2682		0.1497		0.0713	
BSBZV s.e. ($B = 60$):	Mean	0.2053		0.1240		0.0639	
	Median	0.1726		0.1110		0.0586	
	Stability	0.2607		0.1445		0.0716	
BSBZV s.e. ($B = 300$):	Mean	0.1978		0.1210		0.0651	
	Median	0.1702		0.1123		0.0592	
	Stability	0.2607		0.1445		0.0716	

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

Table 10: Standards errors in four variable confirmatory factor analysis model of Figure 2(b), lognormal ζ_2 , incorrectly specified model with zero population variance of δ_1 .

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Heywood case: $\hat{\theta}_1$ (population value $\theta_1 = 0$)							
MC s.d.	0.0634	0.0430	0.0298	0.0190	0.0131	0.0098	0.0060
IM s.e.:	Mean	0.0615	0.0417	0.0297	0.0186	0.0133	0.0094
	Median	0.0589	0.0410	0.0293	0.0185	0.0132	0.0094
	Stability	0.0404	0.0217	0.0101	0.0037	0.0063	0.0013
WLS s.e.:	Mean	0.0507	0.0361	0.0263	0.0166	0.0118	0.0084
	Median	0.0482	0.0353	0.0259	0.0165	0.0117	0.0084
	Stability	0.0603	0.0324	0.0161	0.0065	0.0059	0.0032
ES s.e.:	Mean	0.0640	0.0422	0.0301	0.0187	0.0134	0.0095
	Median	0.0597	0.0411	0.0295	0.0186	0.0133	0.0095
	Stability	0.0707	0.0343	0.0168	0.0067	0.0077	0.0021
EB s.e. ($B = 60$):	Mean	0.0720		0.0303		0.0133	
	Median	0.0640		0.0296		0.0134	
	Stability	0.0633		0.0258		0.0156	
EB s.e. ($B = 300$):	Mean	0.0708		0.0306		0.0134	
	Median	0.0623		0.0301		0.0134	
	Stability	0.0673		0.0184		0.0086	
BSBPF s.e. ($B = 60$):	Mean	0.0583		0.0274		0.0123	
	Median	0.0525		0.0270		0.0123	
	Stability	0.0706		0.0250		0.0119	
BSBPF s.e. ($B = 300$):	Mean	0.0561		0.0277		0.0123	
	Median	0.0515		0.0274		0.0123	
	Stability	0.0573		0.0181		0.0066	
BSBZV s.e. ($B = 60$):	Mean	0.0529		0.0275		0.0122	
	Median	0.0505		0.0269		0.0121	
	Stability	0.0690		0.0220		0.0130	
BSBZV s.e. ($B = 300$):	Mean	0.0544		0.0272		0.0123	
	Median	0.0520		0.0267		0.0123	
	Stability	0.0590		0.0172		0.0056	

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

Table 10: (continued).

Characteristic	Sample sizes						
	50	100	200	500	1000	2000	5000
Non-Heywood case: $\hat{\theta}_3$ (population value $\theta_3 = 0.823$)							
MC s.d.	0.1664	0.1231	0.0845	0.0528	0.0400	0.0259	0.0167
IM s.e.:	Mean	0.1635	0.1166	0.0824	0.0525	0.0373	0.0264
	Median	0.1608	0.1157	0.0822	0.0524	0.0373	0.0264
	Stability	0.1664	0.1231	0.0845	0.0528	0.0400	0.0259
WLS s.e.:	Mean	0.1559	0.1151	0.0829	0.0533	0.0378	0.0269
	Median	0.1522	0.1134	0.0823	0.0531	0.0377	0.0268
	Stability	0.1664	0.1231	0.0846	0.0528	0.0400	0.0259
ES s.e.:	Mean	0.1604	0.1150	0.0819	0.0524	0.0372	0.0264
	Median	0.1552	0.1130	0.0812	0.0522	0.0372	0.0264
	Stability	0.1664	0.1231	0.0845	0.0528	0.0400	0.0259
EB s.e. ($B = 60$):	Mean	0.1610		0.0813		0.0373	
	Median	0.1531		0.0799		0.0371	
	Stability	0.1664		0.0852		0.0398	
EB s.e. ($B = 300$):	Mean	0.1613		0.0815		0.0371	
	Median	0.1554		0.0806		0.0371	
	Stability	0.1663		0.0852		0.0398	
BSBPF s.e. ($B = 60$):	Mean	0.1587		0.0825		0.0378	
	Median	0.1550		0.0815		0.0377	
	Stability	0.1659		0.0852		0.0396	
BSBPF s.e. ($B = 300$):	Mean	0.1540		0.0818		0.0378	
	Median	0.1512		0.0808		0.0377	
	Stability	0.1656		0.0852		0.0396	
BSBZV s.e. ($B = 60$):	Mean	0.1537		0.0823		0.0378	
	Median	0.1491		0.0814		0.0377	
	Stability	0.1606		0.0838		0.0392	
BSBZV s.e. ($B = 300$):	Mean	0.1560		0.0821		0.0380	
	Median	0.1511		0.0809		0.0380	
	Stability	0.1606		0.0838		0.0392	

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to Perfectly Fit the CFA model; BSBZV: Bollen-Stine bootstrap to fit the CFA model with Zero Variance.

appear to reject too often, although this finding is not very firm.

Table 13 describes an interesting situation in which the structural model is misspecified, and the Heywood case is exactly $\theta_1 = 0$ in the population. Since the model is structurally misspecified, WLS standard error perform poorly, with confidence intervals being too narrow, and rejection rates being too high. Bollen-Stine bootstrap percentile test rejects too often with the smallest sample size. All other tests have the required sizes. Since the normal theory inference is not affected by the non-normality of the distribution, it should be concluded that the assumptions of asymptotic robustness are satisfied for this model, contrary to what was expected given its structure.

4.2.3 Power of the tests

As indicated in the last two columns of Table 2, power comparisons are of interest for the tests with known (and identical) sizes.

Changing the values of the parameters in the model of Fig. 2 can create a population Heywood case. For instance, when the coefficient β_{12} of regression of y_2 on y_1 goes down to 0.6, a mild Heywood case appears with $\theta_1 = -0.006$. The asymptotic power of Wald tests can be computed using (46) where the asymptotic variance $\text{As.Var}[\hat{\psi}_1] = 0.509/N$ for data from multivariate normal or asymptotically robust distributions. For the multivariate $t(5)$ data, this variance needs to be increased by a factor of 5/3 to account for heavier tails of the distribution. The theoretical calculations are compared with the simulated power on Fig. 3. In simulations, 2000 Monte Carlo replicates were taken for the analytic results, and 1000 replicates for the bootstrap results. The theoretical power is given by the smooth curve. Different versions of the Wald test are given by circles for analytic standard errors, and by plus signs for Bollen-Stine bootstrap standard errors; the two-sided difference tests, by triangles; the one-sided difference tests, by X's; and the percentile bootstrap tests, by diamonds. The upper panel gives the results for the multivariate normal distribution; the middle panel, for the multivariate $t(5)$ distribution; and the lower panel, for the data with lognormal ζ_2 . Wald tests based on the WLS standard errors were excluded from consideration, as they were biased when the model was misspecified. Also, the performance of the tests with the smallest sample sizes may be problematic, as shown in Tables 11–13. No clear ordering of the tests in terms of their power can be deduced from these results, except that the two-sided tests have lower power, as expected. The bootstrap percentile tests, as well as one-sided difference (signed root) tests might have a slight advantage in small samples, but they were shown in the aforementioned tables to have slightly elevated sizes. The distributionally robust scaled chi-square difference test had somewhat lower power for the non-normal data in Fig. 3(c) than other one-sided tests. The power of all tests is lower in the multivariate $t(5)$ case due to larger sampling variances of the estimates.

The low power of the tests in this example is caused by the population value of the Heywood case that is very close to zero. We tried different sets of parameter values with the data generating model of Fig. 2, and other patterns of the power curves were produced. For instance, with $\beta_{12} = 0.4$, $\beta_{13} = 0.8$, $\beta_{14} = 0.6$, $\beta_{23} = -0.1$, $\beta_{24} = -0.1$, $\mathbb{V}[\delta_3] = \mathbb{V}[\delta_4] = 1$, the obtained Heywood case has a magnitude of $\theta_{01} = -0.1951$. Due to the larger magnitude of the Heywood case (or, to be precise, the ratio of the Heywood case value in population to the asymptotic variance of the corresponding estimate), the tests have non-trivial power that approaches 1 in large samples. While the Wald tests perform poorly with sample sizes

Table 11: Test sizes in four variable confirmatory factor analysis model of Figure 2(b), multivariate normal distribution, incorrectly specified model with zero population variance of δ_1 .

Misspecification test	Sample sizes						
	50	100	200	500	1000	2000	5000
Wald tests							
IM s.e.	5.036	5.106	3.760	4.677	5.072	5.587	5.268
WLS s.e.	6.512	5.568	4.113	4.738	4.994	5.556	5.171
ES s.e.	5.433	5.229	4.034	4.677	4.908	5.493	5.204
EB s.e. ($B = 60$)	4.686		4.518		5.025		
EB s.e. ($B = 300$)	4.289		3.755		5.448		
BSBPF s.e. ($B = 60$)	5.707		4.521		5.532		
BSBPF s.e. ($B = 300$)	5.128		4.064		5.034		
Confidence intervals (coverage)							
IM s.e.	94.964	95.171	94.869	95.723	94.928	94.941	94.215
WLS s.e.	<i>92.609</i>	93.725	94.595	95.508	94.597	95.034	94.215
ES s.e.	94.087	94.432	94.634	95.508	94.519	95.003	94.247
EB s.e. ($B = 60$)	94.678		94.793		93.886		
EB s.e. ($B = 300$)	95.314		94.789		94.216		
BSBPF s.e. ($B = 60$)	93.962		94.559		94.049		
BSBPF s.e. ($B = 300$)	94.210		95.322		94.379		
Bootstrap percentile tests							
EB %-tile ($B = 60$)	7.943		5.666		5.779		
EB %-tile ($B = 300$)	<i>8.261</i>		5.211		5.532		
BSBPF %-tile ($B = 60$)	7.361		5.977		6.203		
BSBPF %-tile ($B = 300$)	7.858		4.985		5.453		
One-sided local tests							
NT test at the boundary	6.782	5.350	3.911	4.750	5.285	5.727	5.181
Scaled test at the boundary	<i>8.354</i>	6.250	4.218	4.750	5.369	5.626	5.181
Signed root of NT difference	6.782	5.350	3.911	4.750	5.285	5.727	5.181
Signed root of scaled difference	<i>8.354</i>	6.250	4.218	4.750	5.369	5.626	5.181
Two-sided local tests							
χ^2 -difference	6.369	6.000	5.215	4.350	5.369	5.170	5.397
Scaled difference	<i>8.271</i>	<i>7.050</i>	5.215	4.700	5.369	5.220	5.505

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to fit the CFA model.

Simulation margin of error: for analytic standard errors and tests, $\sqrt{0.95 \cdot 0.05/2000} = 0.49\%$; for the bootstrap standard errors and tests, $\sqrt{0.95 \cdot 0.05/1000} = 0.69\%$. *Emphasis in italics* indicates test sizes different from 5% or coverage different from 95% by 2% or more for analytic standard errors and tests, and by 3% or more for the bootstrap standard errors and tests.

Table 12: Test sizes in four variable confirmatory factor analysis model of Figure 2(b), multivariate $t(5)$ distribution, incorrectly specified model with zero population variance of δ_1 .

Misspecification test	Sample sizes						
	50	100	200	500	1000	2000	5000
Wald tests							
IM s.e.	<i>9.164</i>	<i>11.333</i>	<i>12.831</i>	<i>13.486</i>	<i>14.180</i>	<i>14.939</i>	<i>15.512</i>
WLS s.e.	<i>7.813</i>	<i>7.145</i>	6.479	6.124	5.302	5.543	5.386
ES s.e.	5.084	5.621	5.255	5.784	5.064	5.262	5.448
EB s.e. ($B = 60$)	2.798		4.896		5.766		
EB s.e. ($B = 300$)	2.238		4.647		5.463		
BSBPF s.e. ($B = 60$)	5.276		5.812		6.110		
BSBPF s.e. ($B = 300$)	4.480		4.962		5.349		
Confidence intervals (coverage)							
IM s.e.	<i>88.195</i>	<i>84.667</i>	<i>83.177</i>	<i>79.957</i>	<i>80.327</i>	<i>79.017</i>	<i>77.624</i>
WLS s.e.	<i>90.905</i>	93.045	94.295	94.618	94.854	95.584	95.137
ES s.e.	94.596	94.443	95.193	94.958	95.053	95.647	95.199
EB s.e. ($B = 60$)	95.923		94.357		95.068		
EB s.e. ($B = 300$)	96.563		95.104		94.537		
BSBPF s.e. ($B = 60$)	93.525		93.675		93.581		
BSBPF s.e. ($B = 300$)	94.640		94.012		94.341		
Bootstrap percentile tests							
EB %-tile ($B = 60$)	7.034		<i>8.050</i>		7.208		
EB %-tile ($B = 300$)	6.075		7.386		6.373		
BSBPF %-tile ($B = 60$)	<i>8.393</i>		<i>8.291</i>		7.270		
BSBPF %-tile ($B = 300$)	8.000		7.528		6.434		
One-sided local tests							
NT test at the boundary	<i>12.320</i>	<i>12.590</i>	<i>12.575</i>	<i>13.939</i>	<i>15.581</i>	<i>14.800</i>	<i>14.850</i>
Scaled test at the boundary	<i>7.520</i>	6.347	4.876	5.808	5.194	4.450	4.900
Signed root of NT difference	<i>12.320</i>	<i>12.590</i>	<i>12.575</i>	<i>13.939</i>	<i>15.581</i>	<i>14.800</i>	<i>14.850</i>
Signed root of scaled difference	<i>7.520</i>	6.347	4.876	5.808	5.194	4.450	4.900
Two-sided local tests							
χ^2 -difference	<i>16.080</i>	<i>15.841</i>	<i>18.306</i>	<i>20.909</i>	<i>20.000</i>	<i>20.400</i>	<i>22.350</i>
Scaled difference ⁹	<i>9.920</i>	<i>8.772</i>	<i>7.613</i>	6.768	6.357	5.250	5.550

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to fit the CFA model; NT: normal theory test statistic.

Simulation margin of error: for analytic standard errors and tests, $\sqrt{0.95 \cdot 0.05/2000} = 0.49\%$; for the bootstrap standard errors and tests, $\sqrt{0.95 \cdot 0.05/1000} = 0.69\%$. *Emphasis in italics* indicates test sizes different from 5% or coverage different from 95% by 2% or more for analytic standard errors and tests, and by 3% or more for the bootstrap standard errors and tests.

⁹Negative values of the scaled difference test statistics were observed in 4 Monte Carlo samples of size $N = 50$ and 2 samples of size $N = 100$.

Table 13: Test sizes in four variable confirmatory factor analysis model of Figure 2(b), lognormal ζ_2 , incorrectly specified model with zero population variance of δ_1 .

Misspecification test	Sample sizes						
	50	100	200	500	1000	2000	5000
Wald tests							
IM s.e.	4.276	4.278	4.933	6.214	4.056	4.045	5.249
WLS s.e.	<i>8.641</i>	6.366	<i>7.121</i>	<i>8.474</i>	6.537	5.931	<i>8.214</i>
ES s.e.	4.274	3.873	4.667	5.817	3.876	3.702	4.976
EB s.e. ($B = 60$)	3.825		4.728		4.351		
EB s.e. ($B = 300$)	2.235		4.247		3.780		
BSBPF s.e. ($B = 60$)	5.071		6.651		5.484		
BSBPF s.e. ($B = 300$)	4.417		6.250		5.489		
Confidence intervals (coverage)							
IM s.e.	96.392	94.611	94.889	94.612	95.493	94.515	94.615
WLS s.e.	<i>89.666</i>	<i>90.502</i>	<i>91.811</i>	<i>91.167</i>	<i>92.651</i>	<i>90.984</i>	<i>91.241</i>
ES s.e.	96.260	94.880	95.333	94.973	95.809	94.789	94.751
EB s.e. ($B = 60$)	96.892		95.192		94.910		
EB s.e. ($B = 300$)	97.526		95.433		95.234		
BSBPF s.e. ($B = 60$)	92.436		92.708		94.259		
BSBPF s.e. ($B = 300$)	92.667		93.029		93.825		
Bootstrap percentile tests							
EB %-tile ($B = 60$)	5.896		5.128		5.337		
EB %-tile ($B = 300$)	5.108		5.449		4.108		
BSBPF %-tile ($B = 60$)	<i>8.645</i>		7.612		6.512		
BSBPF %-tile ($B = 300$)	7.500		6.971		6.089		
One-sided local tests							
NT test at the boundary	5.421	4.990	5.288	6.350	3.774	4.692	5.644
Scaled test at the boundary	5.671	4.788	4.888	5.908	3.431	3.784	4.618
Signed root of NT difference	5.421	4.990	5.288	6.350	3.774	4.692	5.644
Signed root of scaled difference	5.671	4.788	4.888	5.908	3.431	3.784	4.618
Two-sided local tests							
χ^2 -difference	4.337	6.250	5.849	5.743	4.803	6.004	5.695
Scaled difference	5.671	6.351	4.567	4.804	4.031	5.197	4.259

IM: information matrix; WLS: weighted least squares; ES: empirical sandwich; EB: empirical bootstrap; BSBPF: Bollen-Stine bootstrap with data rotated to fit the CFA model; NT: normal theory test statistic.

Simulation margin of error: for analytic standard errors and tests, $\sqrt{0.95 \cdot 0.05/2000} = 0.49\%$; for the bootstrap standard errors and tests, $\sqrt{0.95 \cdot 0.05/1000} = 0.69\%$. *Emphasis in italics* indicates test sizes different from 5% or coverage different from 95% by 2% or more for analytic standard errors and tests, and by 3% or more for the bootstrap standard errors and tests.

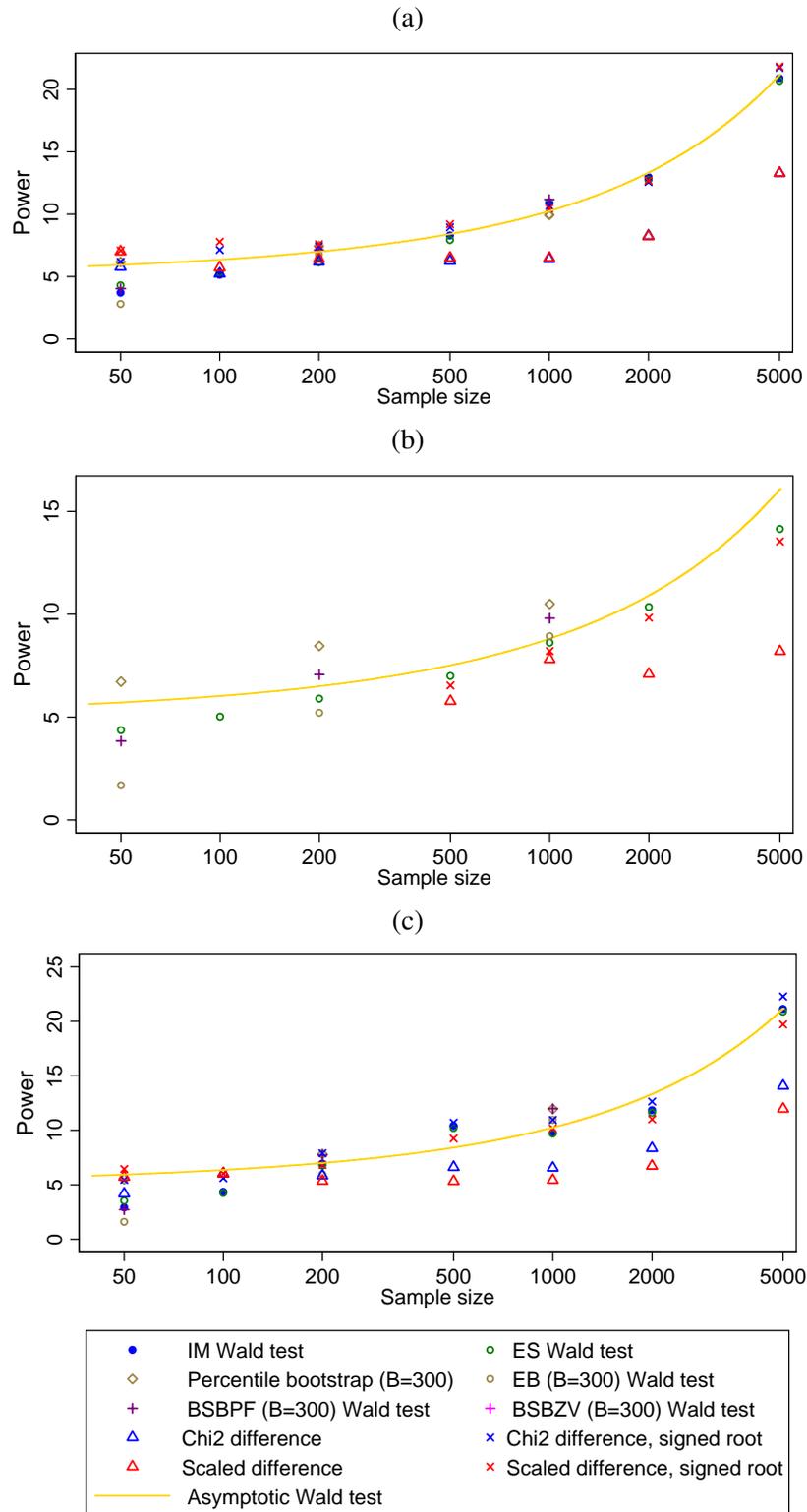


Figure 3: Power curves of the Heywood case tests, Heywood case = -0.006 . (a), multivariate normal data; (b) multivariate $t(5)$ data; (c), data with lognormal ζ_2 .

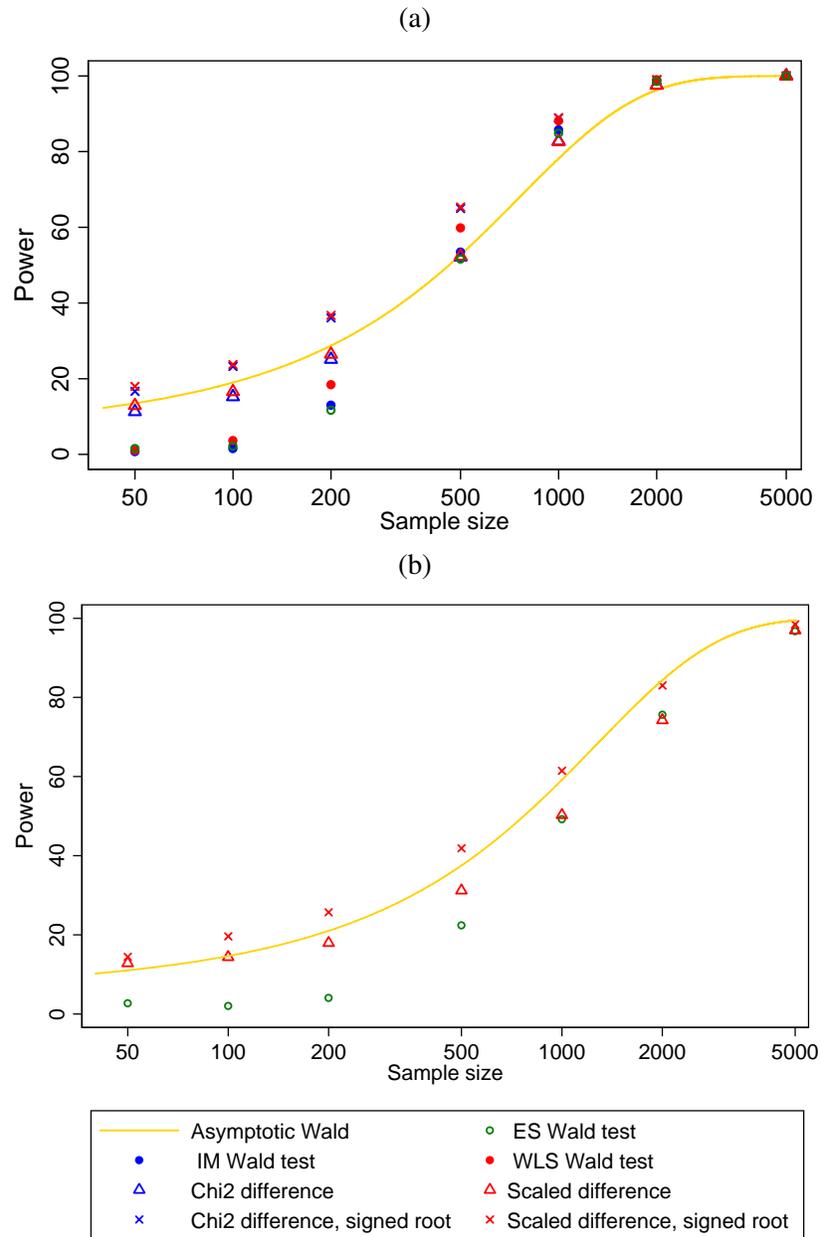


Figure 4: Power curves of the Heywood case tests, Heywood case = -0.195 . (a), multivariate normal data; (b) multivariate $t(5)$ data.

up to $N \leq 500$, the signed root tests outperformed other tests in terms of power. No bootstrap simulations were performed for this setting.

5 Empirical example

The previous section presented results of simulated data where we knew the correct and misspecified models. In this section we illustrate the use of our tests with real empirical data where a negative error variance emerges.

The multiple traits, multiple methods model represents two dimensions of liberal democracy in countries. The first dimension is Political Rights and the second is Political Liberties. Political rights refer to the accountability of elites to nonelites in free and fair elections. Political liberties refer to the freedom of expression and the freedom to organize. Eight indicators come from three judges and the model includes systematic error shared among indicators that come from the same judge (either Sussman method, Gastil method, or Banks method). Details on the indicators and more discussion of the model is in Bollen (1993). The specific variables used in the analysis are given in Table 14, and the path diagram, in Fig. 5.

Table 14: Variables in cross-national liberal democracy example of Bollen (1993).

Variable	Factor 1: political liberalization	Factor 2: democratic rule	Factor 3: Sussman method	Factor 4: Gastil method	Factor 5 : Banks method
Party formation	$\lambda_{11} = 1$				λ_{15}
Broadcast media	λ_{21}		$\lambda_{23} = 1$		
Printed media	λ_{31}		λ_{33}		
Civil liberties	λ_{41}			$\lambda_{41} = 1$	
Legislative efficiency		$\lambda_{52} = 1$			$\lambda_{55} = 1$
Political rights		λ_{62}		λ_{64}	
Competitive nomination		λ_{72}			λ_{75}
Effective selection		λ_{82}			λ_{85}

Table 15 reports the estimation results with several variance estimators considered in the paper. Since the overall fit of the model was not rejected, all types of standard errors except for the normal theory observed information matrix ones can be considered applicable. The naïve standard errors will be expected to be biased since the data are ordinal and severely non-normal (the p -values of Mardia multivariate skewness and kurtosis are 0.000 and 0.003, respectively).

Our particular interest lies in Wald tests for Heywood cases given in the last line of the table. We see that different estimators and tests give a very similar picture, none of them rejecting the null of a non-negative error variance.

The signed root of the likelihood ratio was -0.606 , and the signed root of the Satorra & Bentler (2001) scaled T statistic was -0.609 , both giving one-sided p -value of 0.27. This type of test is known to have greater power, and thus it is not surprising that p -values are somewhat smaller than those of Wald tests.

According to the general asymptotic robustness theory (Anderson & Amemiya 1988, Satorra 1990), if

Table 15: Parameter estimates and standard errors in liberal democracy factor analysis of Bollen (1993).

	Point estimate	Standard error estimation method				
		Normal theory observed info	WLS	Empirical sandwich	Bollen-Stine bootstrap	Empirical bootstrap
Political liberalization factor, $\lambda_{11} = 1$						
λ_{21}	0.8605	0.0592	0.0654	0.0599	0.0634	0.0613
λ_{31}	0.9250	0.0616	0.0579	0.0599	0.0556	0.0507
λ_{41}	0.7188	0.0432	0.0434	0.0470	0.0424	0.0436
Democratic rule factor, $\lambda_{52} = 1$						
λ_{62}	1.0780	0.0598	0.0659	0.0583	0.0586	0.0562
λ_{72}	0.9394	0.0501	0.0597	0.0438	0.0484	0.0477
λ_{82}	0.4380	0.0838	0.0780	0.0742	0.0723	0.0692
Sussman method factor, $\lambda_{23} = 1$						
λ_{33}	1.1912	0.2271	0.2314	0.2872	0.4060	0.3884
Gastil method factor, $\lambda_{44} = 1$						
λ_{64}	0.6328	0.1573	0.1780	0.1985	0.1770	0.1665
Banks method factor, $\lambda_{55} = 1$						
λ_{15}	-0.1836	0.3875	0.6227	0.4410	0.6227	0.4802
λ_{75}	2.7110	0.7401	0.7441	0.8689	1.0454	0.8772
λ_{85}	1.9365	0.5309	0.6182	0.5597	0.7356	0.6056
Factor variances and covariances						
ϕ_{11}	16.0299	2.1239	1.3829	1.4558	1.4773	1.4392
ϕ_{22}	10.4852	1.4833	1.1202	1.1008	1.0996	1.1747
ϕ_{12}	12.8593	1.6150	1.1130	1.0588	1.1053	1.1916
ϕ_{33}	2.5688	0.8175	1.1112	1.0390	1.0048	1.0458
ϕ_{44}	1.4325	0.4455	0.4740	0.5171	0.5146	0.5090
ϕ_{55}	0.6788	0.3843	0.4804	0.4381	0.3998	0.4528
ϕ_{34}	1.4721	0.4716	0.6606	0.5905	0.5517	0.5704
ϕ_{35}	-0.2802	0.2143	0.3093	0.3102	0.1990	0.3006
ϕ_{45}	-0.3428	0.1818	0.2509	0.2272	0.1569	0.2431
Measurement error variances						
θ_1	2.0940	0.6098	0.8900	1.0465	0.7750	0.8581
θ_2	3.0922	0.5004	0.4885	0.5152	0.5156	0.5092
θ_3	1.5723	0.5489	0.5140	0.5768	0.7053	0.7274
θ_4	0.6068	0.1935	0.1927	0.2135	0.3414	0.3151
θ_5	1.5788	0.2788	0.2680	0.3211	0.3042	0.3113
θ_6	0.2689	0.2152	0.3683	0.2592	0.1999	0.2404
θ_7	-0.4186	1.0740	0.8945	1.2389	1.2210	1.3313
θ_8	8.4993	1.1334	1.0681	1.2024	1.0679	1.0949
Goodness of fit		$\chi^2(8) = 9.21$	$\chi^2(7.4) = 8.19$			
<i>p</i> -value		0.33	0.36		0.33	0.38
# bootstrap samples, converged/total					258/300	230/300
Heywood case Wald test $H_0 : \theta_7 \geq 0$, <i>p</i> -value		0.35	0.32	0.37	0.37	0.38

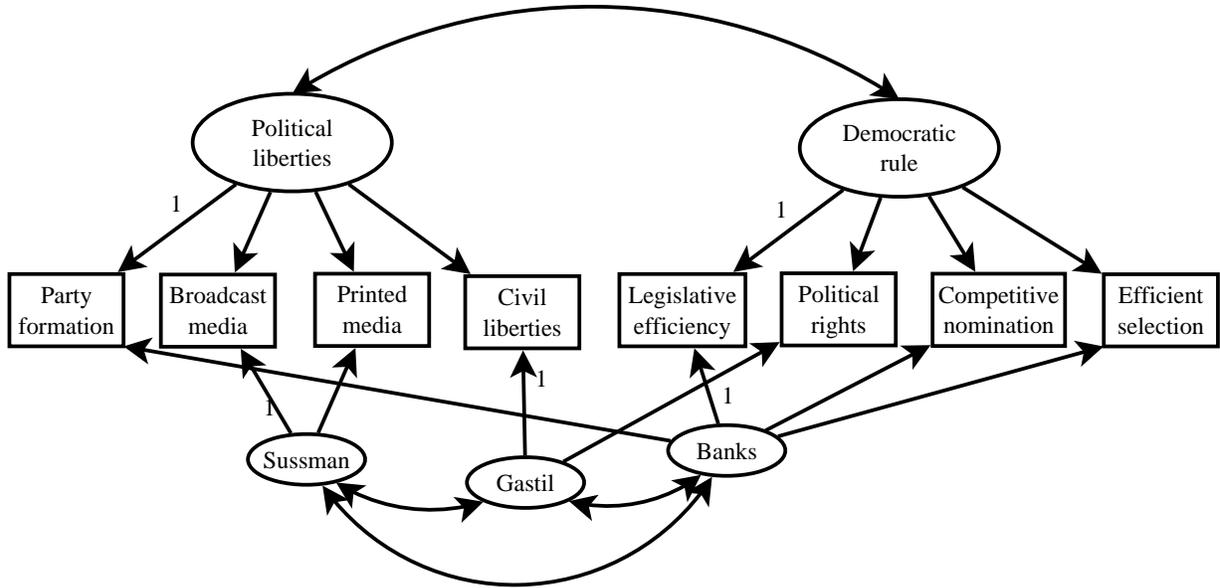


Figure 5: Factor analysis model of cross-national liberal democracy.

the model is specified correctly and the error terms are independent of factors, all factor loadings are efficiently estimated with the inverse of the information matrix providing appropriate standard errors, while variances and covariances of non-normal errors will not be handled properly with the normal theory methods. There is some support to this effect in Table 15. While the standard errors for λ 's do not differ by more than 10% between methods, the standard errors for the factor variances and covariances can differ by some 50% (e.g., estimates of the standard error of ϕ_{11} , which are smaller for the distribution free standard errors, or those of ϕ_{35} , θ_1 and θ_4 which are larger for kurtosis-correcting standard errors).

6 Discussion and Conclusion

We began our paper by explaining that negative variance estimates might occur due to sampling variability, especially with small sample size; underidentification; outliers and influential cases; or structural misspecifications. Diagnostics are available to examine all of these possible causes except that tests to distinguish structural misspecifications versus sampling fluctuations are not well-developed. We assume that the researcher has applied other diagnostics to rule out outliers, influential cases or empirical underidentification as the cause of a Heywood case. Our paper has focused on determining whether the Heywood case resulted from structural misspecification. We proposed new tests and examined these and other tests of Heywood cases to see which worked best under what conditions.

We have reviewed existing proposals that include the Wald tests, confidence intervals, including the bootstrap-based ones, and likelihood ratio tests. We expanded the existing methods to construct Wald tests to include the empirical sandwich estimator that has been used fragmentarily in SEM literature. We also proposed new tests that account for the one-sided nature of the proper vs. improper solutions. We discussed

how the constrained likelihood ratio or, more generally, scaled chi-square difference tests can be used to test Heywood cases. We also introduced signed-root tests based on the likelihood ratio and scaled chi-square difference, and demonstrated their equivalence.

Several factors can potentially influence the performance of tests: the distribution of the observed variables, the sample size, and the structural misspecifications of the model. For the purposes of summarizing our results we distinguish several combinations of these conditions. Prior studies have concentrated on correctly specified structures while we are the first to look at what happens when the Heywood case is due to structural misspecification.

To summarize our recommendations, we shall take as a starting point the possibility that the model is misspecified. This immediately rules out WLS standard errors and the Wald tests based on them. Exactly identified models are exceptions, as in this case $\sigma_* = 0$, and WLS standard errors are the same as the empirical sandwich estimator. The recommendations for other tests can be given based on the breakdown of the data features according to the multivariate kurtosis properties (as an approximation to the more relevant asymptotic robustness which cannot be directly verified) and sample size. Generally, the one-sided chi-square difference (scaled, if needed) tests (either referred to the chi-bar distribution to account for the boundary, or used in a signed root form) and the bootstrap percentile tests work slightly better than Wald tests, especially in smaller samples where the bias of parameter estimates may render Wald tests unreliable. However, the one-sided (scaled) chi-square difference tests had problems in small samples. With non-normal data for which the asymptotic robustness conditions do not hold, empirical sandwich standard errors or empirical bootstrap methods should be employed. If the data demonstrate kurtosis comparable to that of the multivariate normal distribution, or if asymptotic robustness conditions can be verified to hold, the normal theory methods can be used. The latter would include the one-sided chi-square difference tests and information matrix-based standard errors. No harm would be incurred if empirical sandwich standard errors were used with normal data.

The summary of our recommendations for structurally misspecified models is given in Table 16. The practicality of these recommendations is attenuated by lack of the implementation of the proposed procedures in the existing SEM packages. Neither one-sided (scaled) difference tests, nor one-sided empirical bootstrap confidence intervals, nor empirical sandwich variance estimator are readily available in the existing SEM software. The researchers proficient with the general purpose statistical software like Stata or R, however, will have no difficulties in producing the necessary statistics.

Among other findings of interests, the following can be mentioned, in no particular order.

1. Negative variance parameters evidence enormous small sample biases. The magnitude of relative bias may be as large as 100% in the samples of size $N = 50$.
2. The bootstrap standard errors are less stable than the empirical sandwich estimator, unless the number of bootstrap samples exceeds the sample size by a factor of at least 2. The existing software defaults of using a fixed number of bootstrap may be an overkill for small samples, yet insufficient in large samples.
3. There is no “magic” large sample number. For some settings, the asymptotic properties and distributions worked fine with samples as small as $N = 50$, while with other settings (typically, with heavy tailed data), the sample sizes had to be above $N = 2000$ for asymptotic properties to be reliable.

Table 16: Recommended tests of Heywood cases when the structural equation model may have incorrectly specified structure.

Sample size	Test	Multivariate kurtosis	
		Close to normal	Excess kurtosis
Small ($N < 200$)	1-sided chi-square	NT	–
	Bootstrap CI	EB	EB
	Wald test	IM	ES
Large ($N > 1000$)	1-sided chi-square	NT	SD
	Bootstrap CI	EB	EB
	Wald test	IM	ES

Analytic standard errors: IM, observed or expected information matrix;
 ES, empirical sandwich.

Chi-square difference tests: NT, normal theory; SD, Satorra-Bentler scaled difference.

Resampling methods: EB, empirical bootstrap.

Interestingly, a smaller model of Section 4.1 required larger sample sizes for some of the parameters to achieve their asymptotic distributions.

4. The results on Wald tests were generally inconclusive. While in some situations the best performing standard errors resulted in Wald tests that were quite accurate across a range of sample sizes, they required sample sizes in excess of $N = 1000$ in other settings. We listed Wald tests in Table 16 as the last resort when other methods are not feasible.
5. Bollen-Stine bootstrap produces standard errors in between the empirical bootstrap and the WLS ones. They would have slight downward bias, although it did not appear to affect the test sizes or confidence interval coverage. Applied researchers could still use Bollen-Stine bootstrap standard errors if no other methods are available in their software.

Our study has several limitations. As with any simulation studies, the extent to which our findings can be generalized to other models is unclear. One particular limitation is that because of the gross nature of misspecification studied in this paper, the misspecification is “discrete” in the sense that there are no models with degree of misspecification and non-centrality in between the true model and the fitted model. Other frameworks, such as the one used by Chen et al. (2001) and Bollen, Kirby, Curran, Paxton & Chen (2007) with milder and continuously varying degree of misspecification can be used to further study some of the issues raised in the current paper. We also did not study other types of improper solutions, such as those of factor variances below zero and correlations greater than one.

A related limitation is that many different types of structural misspecification that could lead to Heywood cases are possible. We looked at only a few. The degree to which the specific form of structural misspecification makes a difference remains to be determined.

The results of our paper suggest that researchers should look for methods to improve proposed sandwich estimator's finite sample performance. The performance of this estimator has been studied extensively in the context of heteroskedastic regression models (Eicker 1967, White 1980, Carroll et al. 1998, Kauermann & Carroll 2001, Bera, Suprayitno & Premaratne 2002, Hardin 2003) where the small sample biases have been established both analytically and by simulation. This literature relies on the special linear structures and unbiased estimation of the residual variance, which is unlikely to be directly generalizable to SEM.

The version of the information matrix estimator that we used was observed information matrix. While it can be conjectured that the expected information matrix will produce less accurate results under structural misspecification, we do not have solid evidence on this issue.

Even considering these limitations, our research provides several alternative methods to test whether a negative error variance is a symptom of structural misspecification.

References

- Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, MA, USA.
- Anderson, J. C. & Gerbing, D. (1984), 'The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis', *Psychometrika* **49**, 155–173.
- Anderson, T. W. & Amemiya, Y. (1988), 'The asymptotic normal distribution of estimators in factor analysis under general conditions', **16**(2), 759–771.
- Andrews, D. F. (2007), 'Robust likelihood inference for public policy', *The Canadian Journal of Statistics* **35**(3), 341–350.
- Andrews, D. W. K. (1999), 'Estimation when a parameter is on a boundary', *Econometrica* **67**(6), 1341–1383.
- Andrews, D. W. K. (2001), 'Testing when a parameter is on the boundary of the maintained hypothesis', *Econometrica* **69**(3), 683–734. doi:10.1111/1468-0262.00210.
- Arbuckle, J. (1997), *AMOS Users Guide: Version 3.6*, SmallWaters, Chicago.
- Arminger, G. & Schoenberg, R. J. (1989), 'Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models', *Psychometrika* **54**, 409–426.
- Barndorff-Nielsen, O. E. & Cox, D. R. (1994), *Inference and Asymptotics*, Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, New York.
- Bera, A. K., Suprayitno, T. & Premaratne, G. (2002), 'On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators', *Journal of Statistical Planning and Inference* **108**(1–2), 121–136.
- Beran, R. & Srivastava, M. S. (1985), 'Bootstrap tests and confidence regions for functions of a covariance matrix', *The Annals of Statistics* **13**(1), 95–115.

- Berndt, E., Hall, B., Hall, R. & Hausman, J. (1974), 'Estimation and inference in nonlinear structural models', *Annals of Economic and Social Measurement* **3/4**, 653–665.
- Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.
- Bollen, K. (1993), 'Liberal democracy: Validity and method factors in cross-national measures', *American Journal of Political Science* **37**(4), 1207–1230.
- Bollen, K. A. (1987), 'Outliers and improper solutions: A confirmatory factor analysis example', *Sociological Methods and Research* **15**, 375–384.
- Bollen, K. A. (1989), *Structural Equations with Latent Variables*, Wiley, New York.
- Bollen, K. A. (1996a), 'An alternative two stage least squares (2SLS) estimator for latent variable models', *Psychometrika* **61**(1), 109–121.
- Bollen, K. A. (1996b), A limited-information estimator for LISREL models with and without heteroscedastic errors, in G. Marcoulides & R. Schumaker, eds, 'Advanced Structural Equation Modeling Techniques', Erlbaum, Mahwah, NJ, pp. 227–241.
- Bollen, K. A. & Arminger, G. (1991), 'Observational residuals in factor analysis and structural equation models', *Sociological Methodology* **21**, 235–262.
- Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M. & Chen, F. (2007), 'Latent variable models under misspecification: two-stage least squares (2SLS) and maximum likelihood (ML) estimators', *Sociological Methods and Research* **36**(1), 48–86.
- Bollen, K. & Stine, R. (1992), 'Bootstrapping goodness of fit measures in structural equation models', *Sociological Methods and Research* **21**, 205–229.
- Boomsma, A. (1983), *On the Robustness of LISREL (Maximum Likelihood Estimation) Against Small Sample Size and Nonnormality*, Sociometric Research Foundation, Amsterdam, the Netherlands.
- Boomsma, A. & Hoogland, J. J. (2001), 'The robustness of LISREL modeling revisited', *Structural Equation Modeling: Present and Future*.
- Browne, M. W. (1984), 'Asymptotically distribution-free methods for the analysis of the covariance structures', *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.
- Buse, A. (1982), 'The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note', *The American Statistician* **36**(3), 153–157.
- Cadigan, N. G. (1995), 'Local influence in structural equation models', *Structural Equation Modeling* **2**, 13–30.
- Canty, A. J., Davison, A. C., Hinkley, D. V. & Ventura, V. (2006), 'Bootstrap diagnostics and remedies', *The Canadian Journal of Statistics/La revue canadienne de statistique* **34**(1), 5–27.

- Carroll, R. J., Wang, S., Simpson, D. G., Stromberg, A. J. & Ruppert, D. (1998), The sandwich (robust covariance matrix) estimator, Technical report, Department of Statistics, Texas A & M University, College Station, TX. Available at <http://www.stat.tamu.edu/ftp/pub/rjcarroll/sandwich.pdf>.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. & Kirby, J. (2001), 'Improper solutions in structural equation models: Causes, consequences, and strategies', *Sociological Methods and Research* **25**, 223–251.
- Chernoff, H. (1954), 'On the distribution of the likelihood ratio', *The Annals of Mathematical Statistics* **25**(3), 573–578.
- Davidson, R. & MacKinnon, J. (1993), *Estimation and Inference in Econometrics*, Oxford University Press, New York.
- Dillon, W. R., Kumar, A. & Mulani, N. (1987), 'Offending estimates in covariance structure analysis: Comments on the causes and solutions to Heywood cases', *Psychological Bulletin* **101**, 126–135.
- Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics* **7**, 1–26.
- Efron, B. & Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Chapman & Hall/CRC.
- Eicker, F. (1967), Limit theorems for regressions with unequal and dependent errors, in 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 59–82.
- Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Texts in Statistical Science, Chapman & Hall/CRC, London.
- Fletcher, R. (1980), *Practical Methods of Optimization*, Wiley Interscience, New York.
- Gonzalez, R. & Griffin, D. (2001), 'Testing parameters in structural equation modeling: Every "one" matters.', *Psychological Methods* **6**(3), 258–269.
- Gould, W., Pittblado, J. & Sribney, W. (2006), *Maximum Likelihood Estimation with Stata*, 3rd edn, Stata Press, College Station, TX.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics, Springer, New York.
- Hardin, J. W. (2003), The sandwich estimator of variance, in T. B. Fomby & R. C. Hill, eds, 'Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later', Elsevier, New York.
- Heywood, H. B. (1931), 'On finite sequences of real numbers', *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **134**(824), 486–501.
- Huber, P. (1967), The behavior of the maximum likelihood estimates under nonstandard conditions, in 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 221–233.
- Huber, P. (1974), *Robust Statistics*, Wiley, New York.

- Jennrich, R. (2008), 'Nonparametric estimation of standard errors in covariance analysis using the infinitesimal jackknife', *Psychometrika* **73**(4), 579–594.
- Jöreskog, K. G. (1978), 'Structural analysis of covariance and correlation matrices', *Psychometrika* **43**, 443–477.
- Kauermann, G. & Carroll, R. J. (2001), 'A note on the efficiency of sandwich covariance matrix estimation', *Journal of the American Statistical Association* **96**(456), 1387–1396.
- Kolenikov, S. (2009), 'Confirmatory factor analysis with `confa`', *Stata Journal* **9**(3), 329–373.
- Kovar, J. G., Rao, J. N. K. & Wu, C. F. J. (1988), 'Bootstrap and other methods to measure errors in survey estimates', *Canadian Journal of Statistics* **16**, 25–45.
- Kúdo, A. (1963), 'A multivariate analogue of the one-sided test', *Biometrika* **50**(3 and 4), 403–418.
- Lafontaine, F. & White, K. J. (1986), 'Obtaining any Wald statistic you want', *Economics Letters* **21**, 35–40.
- Lee, S. & Jennrich, R. (1979), 'A study of algorithms for covariance structure analysis with specific comparisons using factor analysis', *Psychometrika* **44**(1), 99–113.
- Magnus, J. R. & Neudecker, H. (1999), *Matrix calculus with applications in statistics and econometrics*, 2nd edn, John Wiley & Sons.
- Moustaki, I. & Victoria-Feser, M.-P. (2006), 'Bounded influence robust estimation in generalized linear latent variable models', *Journal of the American Statistical Association* **101**(474), 644–653. DOI 10.1198/016214505000001320.
- Muthén, B. O. & Satorra, A. (1995), 'Complex sample data in structural equation modeling', *Sociological Methodology* **25**, 267–316.
- Neudecker, H. & Satorra, A. (1991), 'Linear structural relations: Gradient and Hessian of the fitting function', *Statistics and Probability Letters* **11**, 57–61.
- Perlman, M. D. (1969), 'One-sided testing problems in multivariate analysis', *The Annals of Mathematical Statistics* **40**(2), 549–567.
- Phillips, P. C. B. & Park, J. Y. (1988), 'On the formulation of Wald tests of nonlinear restrictions', *Econometrica* **56**, 1065–1083.
- Rabe-Hesketh, S. & Skrondal, A. (2005), *Multilevel and Longitudinal Modeling Using Stata*, Stata Press, College Station, TX.
- Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004), 'Generalized multilevel structural equation modeling', *Psychometrika* **69**(2), 167–190.
- Rindskopf, D. (1984), 'Structural equation models: Empirical identification, Heywood cases, and related problems', *Sociological Methods and Research* **13**, 109–119.

- Sato, M. (1987), 'Pragmatic treatment of improper solutions in factor analysis', *Annals of the Institute of Statistics and Mathematics, part B* **39**, 443–455.
- Satorra, A. (1990), 'Robustness issues in structural equation modeling: A review of recent developments', *Quality and Quantity* **24**, 367–386.
- Satorra, A. (1992), 'Asymptotic robust inference in the analysis of mean and covariance structures', *Sociological Methodology* **22**, 249–278.
- Satorra, A. & Bentler, P. (2001), 'A scaled difference chi-square test statistic for moment structure analysis', *Psychometrika* **66**(4), 507–514.
- Satorra, A. & Bentler, P. M. (1990), 'Model conditions for asymptotic robustness in the analysis of linear relations', *Computational Statistics and Data Analysis* **10**(3), 235–249. doi:10.1016/0167-9473(90)90004-2.
- Satorra, A. & Bentler, P. M. (1994), Corrections to test statistics and standard errors in covariance structure analysis, in A. von Eye & C. C. Clogg, eds, 'Latent variables analysis', Sage, Thousands Oaks, CA, pp. 399–419.
- Savalei, V. (forthcoming), 'Expected vs. observed information in SEM with incomplete normal and nonnormal data', *Psychological Methods*.
- Savalei, V. & Kolenikov, S. (2008), 'Constrained vs. unconstrained estimation in structural equation modeling', *Psychological Methods* **13**, 150–170.
- Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.
- Shapiro, A. (1985), 'Asymptotic distribution of test statistic in the analysis of moment structures under inequality constraints', *Biometrika* **72**(1), 133–144.
- Shapiro, A. (1988), 'Towards a unified theory of inequality constrained testing in multivariate analysis', *International Statistical Review* **56**(1), 49–62.
- Skinner, C. J. (1989), Domain means, regression and multivariate analysis, in C. J. Skinner, D. Holt & T. M. Smith, eds, 'Analysis of Complex Surveys', Wiley, New York, chapter 3, pp. 59–88.
- Skrondal, A. (2000), 'Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom', *Multivariate Behavioral Research* **35**(2), 137–167.
- Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling*, Chapman and Hall/CRC, Boca Raton, Florida.
- Stata Corp. (2007), *Stata Statistical Software: Release 10*, College Station, TX, USA.
- Stoel, R. D., Garre, F. G., Dolan, C. & van den Wittenboer, G. (2006), 'On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints', *Psychological Methods* **11**(4).

- van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.
- Van Driel, O. P. (1978), 'On various causes of improper solutions in maximum likelihood factor analysis', *Psychometrika* **43**, 225–43.
- West, K. D. & Newey, W. K. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.
- White, H. (1980), 'A heteroskedasticity-consistent covariance-matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.
- White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**(1), 1–26.
- White, H. (1996), *Estimation, Inference and Specification Analysis*, Vol. 22 of *Econometric Society Monographs*, Cambridge University Press, New York.
- Yuan, K.-H. & Bentler, P. M. (1997), 'Mean and covariance structure analysis: Theoretical and practical improvements', *Journal of the American Statistical Association* **92**(438), 767–774.
- Yuan, K.-H. & Bentler, P. M. (2007), Structural equation modeling, in C. Rao & S. Sinharay, eds, 'Handbook of Statistics: Psychometrics', Vol. 26 of *Handbook of Statistics*, Elsevier, chapter 10.
- Yuan, K.-H. & Chan, W. (2005), 'On nonequivalence of several procedures of structural equation modeling', *Psychometrika* **70**(4), 791–798.
- Yuan, K.-H. & Hayashi, K. (2006), 'Standard errors in covariance structure models: Asymptotics versus bootstrap', *British Journal of Mathematical and Statistical Psychology* **59**, 397–417.

Appendices

A Numeric maximization, derivatives, and the sandwich estimator

This appendix provides basic information about numeric likelihood maximization procedures, the numeric derivatives that are obtained as a part of that procedure, and combining those derivatives into the sandwich estimator. Exposition is based on Gould et al. (2006), using their notation, and is oriented at implementation in Stata software (Stata Corp. 2007).

While analytic derivatives tend to lead to faster and more accurate algorithms, the numeric derivatives are very general and require no derivations and specialized coding, and they are easier to generalize for other estimation problems, such as complex survey designs (Muthén & Satorra 1995). This approach is closer in spirit to GLLAMM estimation procedures (Rabe-Hesketh et al. 2004, Skrondal & Rabe-Hesketh 2004, Rabe-Hesketh & Skrondal 2005) than to traditional LISREL-type algorithms relying on manipulation of the moment matrices. Both approaches use (24) (or equivalently (10a) and (10b) of Yuan & Hayashi (2006)), but provide different implementations of the resulting estimators. Yuan & Hayashi (2006) take this estimator as a starting point, and further pursue the necessary first and second derivatives of the likelihood

to obtain explicit matrix expressions that would involve the estimated parameters and their functions (such as implied moment matrices and their derivatives), as well as the third and the fourth moments of the data.

The usual preference for analytical computations may not always be warranted. For larger models, computations of the fourth order moments matrix in either Satorra-Bentler or Yuan-Hayashi estimators will require $O(np^4)$ operations, with additional slow element-by-element operations for the derivatives of the moment conditions. The computations of the matrices \hat{A}, \hat{B} in the empirical version of the sandwich estimator will require $O(nt^2)$ operations where t is the number of parameters to be estimated. If the number of parameters is roughly proportional to the number of observed variables (and in most SEMs, each observed variable would require at least one loading and the variance parameter), the analytical derivatives might take longer to compute.

Suppose the sample observations X_i are i.i.d. coming from distribution $f(x, \theta)$ where $f(\cdot)$ is the density (or, in some discrete problems, the probability density function). Then the maximum likelihood estimation problem consists of finding the solution of

$$\ln l(\theta, X) = \sum_{i=1}^n \ln f(x_i, \theta) \rightarrow \max_{\theta} \quad (53)$$

The typical SEM example combines assumptions about the model structure to obtain the parameterized mean and covariance matrices as in (5) with the assumption of multivariate normality to obtain the likelihood. In the common situation that the above problem cannot be solved analytically, an iterative maximization algorithm has to be set up. The most popular class of algorithms are those based on Newton-Raphson iterations:

1. Start with the initial values of θ_1 , obtained from a random search, another estimation routine, supplied by user, etc.
2. Compute the derivatives of the objective function:

$$\begin{aligned} \mathbf{g}(\theta) &= \frac{\partial}{\partial \theta} \ln l(\theta, X) \\ \mathbf{H}(\theta) &= \frac{\partial^2}{\partial \theta \partial \theta'} \ln l(\theta, X) \end{aligned} \quad (54)$$

3. Determine the direction of the next step $\mathbf{d} = -[\mathbf{H}(\theta_k)]^{-1} \mathbf{g}(\theta_k)$
4. Compute the next iteration

$$\theta_{k+1} = \theta_k + h_k \mathbf{d} \quad (55)$$

where the step size h_k is chosen to make an improvement of the objective function along direction \mathbf{d} .

5. Check convergence criteria (usually of the form $\|\mathbf{g}(\theta_k)\| < \epsilon$ for some small $\epsilon \sim 10^{-6}$); if satisfied, stop; if not, return to step 2.

The above algorithm needs the first and second derivatives, $\mathbf{g}(\boldsymbol{\theta}_k)$ and $\mathbf{H}(\boldsymbol{\theta}_k)$. If those are not available from the likelihood program, Stata can compute them numerically. The derivatives are approximated as

$$f'(z) \approx \frac{f(z + \frac{d}{2}) - f(z - \frac{d}{2})}{d},$$

$$f''(z) \approx \frac{f(z + d) - f(z) + f(z - d) - f(z)}{d^2}, \quad (56)$$

(57)

where the step size d is computed so that $f(z)$ and $f(z \pm d)$ differ in about half their digits. Thus if f , the log-likelihood function, is computed in double arithmetics with 16 digits, the differences in both numerators will be computed with accuracy of 8 digits.

There is a number of additional smart features of Stata's likelihood maximizer. If matrix $\mathbf{H}(\boldsymbol{\theta}_k)$ is deemed non-invertible, it uses the steepest ascent algorithm where the update is performed as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - h_k \mathbf{g}(\boldsymbol{\theta}_k) \quad (58)$$

It can take a compromise between the two algorithms by scaling the diagonal elements of $\mathbf{H}(\boldsymbol{\theta}_k)$ to make it invertible, and it can take a spectral decomposition of $\mathbf{H}(\boldsymbol{\theta}_k)$ to identify the subspaces where the (projection of the) likelihood is convex, and thus Newton-Raphson steps can be taken, and the orthogonal subspace where the (projection of the) likelihood is flat, and steepest ascent needs to be performed. Other variations of the maximizer are choices between several options of avoiding direct computation of the $\mathbf{H}(\boldsymbol{\theta}_k)$ matrix: BHHH algorithm (Berndt, Hall, Hall & Hausman 1974) that replaces $\mathbf{H}(\boldsymbol{\theta}_k)$ by $-\mathbf{g}(\boldsymbol{\theta}_k)\mathbf{g}(\boldsymbol{\theta}_k)'$; and DFP and BFGS algorithms that build up improved approximations to $\mathbf{H}(\boldsymbol{\theta}_k)$ (Fletcher 1980).

It is easy to see that the above algorithms provide all the building blocks for the sandwich estimator (24). If the observation-level contributions to those matrices

$$\mathbf{g}_i(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln l(\boldsymbol{\theta}, x_i)$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ln l(\boldsymbol{\theta}, x_i)$$

are provided analytically or computed numerically, then the components of the sandwich estimator are obtained as

$$\hat{A} = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i(\hat{\boldsymbol{\theta}}),$$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\hat{\boldsymbol{\theta}})\mathbf{g}_i(\hat{\boldsymbol{\theta}})' \quad (59)$$

It is also automatically available with Stata's `robust` option, implemented through the lower level `_robust` command that provides the computations in (59). The simplicity of implementation allows relatively straightforward extensions to more complicated situations, such as clustered data and data coming from complex surveys. In those situations, the summations needs to be taken over clusters, or PSUs, and the individual observations might need to be weighted according to sampling weights.