Residuals and the residual-based statistic for testing goodness of fit of structural equation models

Njål Foldnes
BI Norwegian Business School

Tron Foss
BI Norwegian Business School

Ulf Henning Olsson
BI Norwegian Business School

This is the authors' final, accepted and refereed manuscript to the article published in

***Journal of Educational and Behavioral Statistics*, 37(2012)3: 367-386**

DOI: http://dx.doi.org/10.3102/1076998611411920

Residuals and the residual-based statistic for testing goodness of fit of structural equation models

Abstract

The residuals obtained from fitting a structural equation model are crucial ingredients in obtaining chi-square goodness-of-fit statistics for the model. We present a didactic discussion of the residuals, obtaining a geometrical interpretation by recognizing the residuals as the result of oblique projections. This sheds light on the concept of degrees of freedom of the model. We use a simple example to illustrate the theory and also to provide simulations of residuals in three dimensions. We then explain the rationale behind the formula for the residual-based test statistic. The formula for the statistic is deduced using linear algebra and large-sample theory. Details are provided so that this material can be used in graduate instruction.

Keywords: Goodness-of-fit, residuals, degrees of freedom, residual-based statistic

1 Introduction

Given a proposed covariance structure model a basic question that needs to be answered is: Does the model fit the data that we observe? There are various competing ways to measure the goodness of fit of a model, and most of them are based on the discrepancies between observed values and the values predicted under the proposed model. Such discrepancies between observed and estimated values are called residuals:

$$\text{residual} = \text{observation} - \text{fitted value}.$$

For covariance structure models, the observations are the covariances and variances of the observed variables. Various versions of the residual sum of squares give rise to competing χ^2 measures of model fit.

In the first part of this article we study the residuals from a geometric point of view. The residual vector is shown to be the result of projecting the observed vector onto a subspace. In other words, estimation of the model constrains the residuals to live in a linear subspace. The dimension

of this subspace represents the degrees of freedom of the model, giving an interesting interpretation of this concept. These results are quite general and are valid for all consistent estimation methods like unweighted and generalized least squares (ULS and GLS) or normal-theory based maximum likelihood (ML) estimation. Notation and definitions are illustrated with the use of a simple example that is used throughout this paper. This example is a toy model which is far too small to be anything close to what a substantive researcher may use, and it is intended solely for instructional purposes. The smallness of the model ensures that there are only three residuals, two variances and one covariance. This allows us to visualize the residuals in three dimensions, and simulated residuals gives us a visual confirmation of the projection theory.

In the second part we give a didactic presentation of an important but relatively unknown type of χ^2 goodness-of-fit statistics in structural equation modeling (SEM), namely the residual-based statistic T_B introduced by Browne (1984). The residual-based test statistic is not as well known as the minimum fit function (MFF) value statistic obtained by multiplying the minimum fit function value by the number of cases minus one. The most prominent MFF statistic is the normal theory maximum likelihood (ML) statistic T_{ML}. The MFF test statistic is asymptotically distributed as a chi-square provided the data at hand meets the distributional assumptions, e.g. normality, of the estimation method. However, in situations where the estimation method is not correctly specified for the data, the MFF statistic may not be asymptotically distributed as a chi-square distribution. For instance, when data are not normally distributed, T_{ML} will most likely not approximate a chi-square distribution, even for large sample sizes (see Yuan, Bentler, and Zhang (2005) for a clear presentation of the univariate case). In contrast, the residual-based statistic T_B can be used in conjunction with the ML estimates, and it will approximate a chi-square distribution even for non-normal data, for sufficiently large sample sizes. That is, an important application of residual-based tests is in situations where non-optimal estimators have been used and a test statistic with a known (asymptotic) distribution is required. See e.g., Savalei and Bentler (2009); Cai and Lee (2009) for recent examples of the utility of residual-based test statistics in a two stage procedure designed to handle missing data. The mathematically inclined reader may consult Shapiro (2007) for a thorough tutorial on statistical inference in covariance structure analysis. A comprehensive overview of estimation methods and test statistics for mean and covariance structures can be found in Yuan and Bentler (2007).

Residual-based test statistics are routinely used to evaluate whether a model is valid or not. However, the formula for the residual-based statistic is quite complicated. In our experience many students and researchers have difficulties in understanding the formula, as it is given by a matrix algebra expression and involves linear algebraic concepts. Our aim is to work out the construction of T_B in detail and explain how it is used to test the fit of the proposed model. In this we broadly follow the seminal

work in Browne (1984).

In the following we introduce notation and definitions. Suppose $\boldsymbol{x}$ is a stochastic p-vector of observed variables with population covariance matrix $\boldsymbol{\Sigma}$. Let the free parameters in the proposed model be contained in the q-vector $\boldsymbol{\theta}$. A structural equation model then implies a certain parametrization $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of the covariance matrix of the observed variables. The null hypothesis states that the model is correctly specified, meaning that there are parameter values such that the model-implied covariance matrix equals the population covariance matrix. This is written as

$$H_0 : \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} \quad \text{for some } \boldsymbol{\theta}.$$

In other words, we say that the model holds if there exists a parameter value $\boldsymbol{\theta}_0$ such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}$. In the following we assume that $\boldsymbol{\theta}_0$ is unique, i.e. that the model is identified. We also assume that the function $\boldsymbol{\Sigma}(.)$ is continuously differentiable.

Since the sample covariance matrix $\boldsymbol{S}$ and the model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are symmetric, the elements below the diagonal in these matrices are duplicates of elements above the diagonal. A more economical way to work $\boldsymbol{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is to restrict attention to only the non-redundant elements. This is done by forming a column vector from the elements above and including the diagonal taken columwise. If $\boldsymbol{A}$ is a $p \times p$ symmetric matrix, there are $p^* = p(p+1)/2$ such non-rendundant elements. Let $vech(\cdot)$ denote this operator that transforms the matrix $\boldsymbol{A}$ into a p^*-vector $vech(\boldsymbol{A})$. Now we define $\boldsymbol{\sigma}(\boldsymbol{\theta}) = vech(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ and $\boldsymbol{s} = vech(\boldsymbol{S})$ and note that $\boldsymbol{\sigma}(\boldsymbol{\theta})$ and $\boldsymbol{s}$ are both p^*-vectors.

To exemplify the general notation and theory covered in this article, let us introduce a very simple model for didactic purposes.

Figure 1 here.

Example. *Consider the factor model whose path diagram is given in figure* (1)*. The observed variables are contained in the* 2*-vector* $x = (x_1, x_2)'$*. The model specifies that* x *can be regressed upon a single latent variable (factor)* F*. The structural equations are*

$$\begin{aligned} x_1 &= \lambda F + \delta_1 \\ x_2 &= \lambda F + \delta_2 \end{aligned} \tag{1}$$

where the factor loadings are identical. In this model we assume that $var(F) = 1$*,* $cov(F, \delta_i) = 0$ *and* $var(\delta_i) = 1$ *for* $i = 1, 2$ *and that* $cov(\delta_1, \delta_2) = 0$*. Hence our model contains only one free parameter, namely the factor loading* λ*, and we have* $p = 2$*,* $p^* = 2 \cdot 3/2 = 3$ *and* $q = 1$*. As the reader may verify using basic covariace algebra, the model-implied*

covariance matrix and its reduced vector form are given by

$$\boldsymbol{\Sigma}(\lambda) = \begin{pmatrix} \lambda^2 + 1 & \lambda^2 \\ \lambda^2 & \lambda^2 + 1 \end{pmatrix}$$

$$\boldsymbol{\sigma}(\lambda) = vech\left(\boldsymbol{\Sigma}(\lambda)\right) = \begin{pmatrix} \lambda^2 + 1 \\ \lambda^2 \\ \lambda^2 + 1 \end{pmatrix}. \tag{2}$$

This paper is organized as follows. In the first part we study the residuals in covariance structure analysis. Next we simulate the residuals with finite samples for a very small model, and obtain visual confirmation of the residual theory. Next we use the theory to construct the residual-based statistic T_B, before we round off with concluding remarks.

2 The residuals

In this part we give a general treatment of the asymptotic behavior of the residuals. The results are valid for all consistent estimation methods.

The asymptotic distribution of the sample covariance matrix

A central element in estimating and testing a model is the covariance matrix of the observed variables. We are therefore interested in assessing the sampling distribution of $\boldsymbol{s}$. In many situations the finite-sample distribution of $\boldsymbol{s}$ is not known, but may be approximated by considering what happens as $n \to \infty$. As the sample size increases the stochastic vector $\boldsymbol{s}$ converges in probability to the population vector $\boldsymbol{\sigma} = vech(\boldsymbol{\Sigma})$:

$$\boldsymbol{s} \quad \xrightarrow{P} \quad \boldsymbol{\sigma},$$

where $\xrightarrow{P}$ denotes convergence in probability. Informally this means that for large sample sizes $\boldsymbol{s}$ is almost certainly almost equal to $\boldsymbol{\sigma}$. Hence for infinite sample size the random nature of $\boldsymbol{s}$ vanishes and it converges toward the constant $\boldsymbol{\sigma}$. However, by magnifying $\boldsymbol{s}$ by a factor $\sqrt{n}$ the resulting vector has a non-degenerate limiting distribution. That the factor $\sqrt{n}$ is of right size can be seen by noting that the variance of $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma})$ is independent of n. In more technical terms it follows from the multivariate central limit theorem (e.g., Anderson, 2003, Theorem 3.4.3) that

$$\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}) \quad \xrightarrow{d} \quad N(\mathbf{0}, \boldsymbol{\Gamma}). \tag{3}$$

The symbol $\xrightarrow{d}$ denotes convergence in distribution. In other words, expression (3) states that in infinite samples the product $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma})$ follows a normal distribution. For a thorough treatment of asymptotic statistics in general, the reader may consult Vaart (2000), while Satorra (1989) contains a self contained but mathematically advanced review of asymptotic theory for test statistics in SEM.

The asymptotic covariance matrix $\boldsymbol{\Gamma}$ in (3) is assumed to be non-singular. This matrix holds crucial information about the asymptotic distribution of $\boldsymbol{s}$ and it is central in designing well-behaved estimators and goodness-of-fit tests in SEM. Software packages in SEM calculate estimates of $\boldsymbol{\Gamma}$ based on the raw data as a necessary ingredient for robust inferences. If the observable vector $\boldsymbol{x}$ is normally distributed, to calculate $\boldsymbol{\Gamma}$ one can use the following well-known formula: $\boldsymbol{\Gamma = 2K_p'(\Sigma \otimes \Sigma)K_p}$. Here the matrix $\boldsymbol{K}_p$ is a $p^2 \times p^*$ matrix with elements 0, $\frac{1}{2}$ or 1 as shown in Section 2 in Browne (1974). For more about the matrix $\boldsymbol{K}_p$ and related matrices, see p. 46 in Magnus and Neudecker (1999).

Example (continued)**.** *In the previous section we introduced a simple factor model example with two observable variables contained in the* 2*-vector* $\boldsymbol{x} = (x_1, x_2)'$*. We will assume that* $\boldsymbol{x}$ *is the product of the following data-generating process:*

$$\begin{aligned} x_1 &= F + \delta_1 \\ x_2 &= F + \delta_2 \end{aligned} \tag{4}$$

where the random variables F*,* δ_1 *and* δ_2 *are i.i.d. standard normal variables. The reader may verify that this implies that* $\boldsymbol{x}$ *has the following population covariance matrix:*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \textit{and hence} \quad \boldsymbol{\sigma} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}.$$

Comparing (4) *with* (1) *it is clear that the model is correctly specified. To be precise, by setting the free parameter* λ *in the model to* $\lambda_0 = 1$ *the model-implied covariance matrix in* (2) *equals the population covariance matrix above:* $\boldsymbol{\sigma}(1) = \boldsymbol{\sigma}$*.*

The formula $\boldsymbol{\Gamma = 2K_p'(\Sigma \otimes \Sigma)K_p}$ *applied here yields*

$$\begin{aligned} \boldsymbol{\Gamma} &= 2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \left(\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \otimes \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 8 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 8 \end{pmatrix}, \end{aligned}$$

and we have the following version of (3)*, where* s_{ij} *denotes the sample covariance between* x_i *and* x_j*:*

$$\sqrt{n} \left[\begin{pmatrix} s_{11} \\ s_{12} \\ s_{22} \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \right] \quad \xrightarrow{d} \quad N\left(\mathbf{0}, \begin{pmatrix} 8 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 8 \end{pmatrix} \right).$$

Minimum distance estimation

An intuitive way of estimating the population parameters $\boldsymbol{\theta}_0$ is to somehow minimize the distance between the observed covariances $\boldsymbol{s}$ and the model-implied covariances $\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}_n)$. The minimum distance (MD) estimator $\hat{\boldsymbol{\theta}}_n$ of

$\boldsymbol{\theta}_0$ is defined as the minimizer of the quadratic form

$$F(\boldsymbol{\theta} \mid \boldsymbol{V}_n) = (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' \boldsymbol{V}_n (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \tag{5}$$

where $\boldsymbol{V}_n$ converges in probability to a positive definite matrix $\boldsymbol{V}$. In most cases $\boldsymbol{V}_n$ is a stocastic matrix that is evaluated on the basis of the sample at hand. We follow Satorra (2003) and use the term 'minimum distance' for the discrepancy function in (5). Other authors (e.g., Shapiro, 2007) refer to this function as a generalized least squares discrepancy function.

Most estimation methods in current use for covariance structure analysis are MD estimators. As shown in Browne (1974), even maximum likelihood estimation can be thought of as MD estimation. Browne (1984) later showed that any MD estimator is consistent:

$$\hat{\boldsymbol{\theta}}_n \quad \xrightarrow{P} \quad \boldsymbol{\theta}_0, \tag{6}$$

and that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is multivariate normal with zero mean vector. To obtain MD estimators which have minimal standard errors we need to be careful about the choice of the weight matrix $\boldsymbol{V}_n$. We say that the MD estimator is *correctly specified* for the data at hand if

$$\boldsymbol{V}_n \quad \xrightarrow{P} \quad \boldsymbol{\Gamma}^{-1}. \tag{7}$$

This condition ensures that the estimator is asymptotically efficient, meaning that the asymptotic covariance matrix of the estimator attains its lower bound within the class of MD estimators.

Table 1 here.

In Table 1 the weight matrix $\boldsymbol{V}_n$ associated with some common estimation methods is listed. The matrix $\boldsymbol{K}_p^-$ is a left inverse of $\boldsymbol{K}_p$. Note that unweighed least squares estimation (ULS) does not satisfy property (7), since $\boldsymbol{V}_n$ is a constant in ULS estimation. This implies that ULS estimates are not asymptotically optimal, in the sense that for infinite sample size there are other estimators with lower standard errors than the ULS estimator. However, as we shall see, the ULS estimator does not impair the asymptotic (infinite sample) behaviour of the residual-based statistic for testing goodness-of-fit compared to other asymptotically optimal estimators. Provided that the data are multivariate normally distributed general least squares (GLS) and maximum likelihood (ML) estimation satisfy property (7) and are asymptotically optimal in the sense of having minimum standard errors. The estimator $\hat{\boldsymbol{\theta}}$ in the ML estimator weight matrix is the minimizer of the likelihood function.

The weight matrix $\widehat{\boldsymbol{A}}$ used in the asymptotically distribution-free (ADF) estimation method of Browne (1984) involves calculating fourth-order central sample moments. $\widehat{\boldsymbol{A}}^{-1}$ satisfies property (7) for the wide range of distributions with finite fourth-order moments. But although consistent for a variety of distributions of the data, $\widehat{\boldsymbol{A}}^{-1}$ has a slow rate of convergence. The high variability of the ADF estimator renders it useful only for medium to large sample sizes.

The asymptotic distribution of MD estimators

For finite samples the distribution of the MD estimator $\hat{\boldsymbol{\theta}}_n$ is very difficult to calculate exactly. However, we shall see that the MD estimator is asymptotically normally distributed.

A central matrix is the the Jacobian matrix of partial derivatives of the function $\boldsymbol{\sigma}(\boldsymbol{\theta})$, i.e. the $p^* \times q$ matrix

$$\boldsymbol{\Delta}(\boldsymbol{\theta}) \equiv \left(\frac{\partial \boldsymbol{\sigma}_i(\boldsymbol{\theta})}{\partial \theta_j} \right)_{i \leq p^*,\, j \leq q}$$

The notation "$\equiv$" means "equal by definition". Note that $\boldsymbol{\Delta}(\boldsymbol{\theta})$ can be evaluated at different values of the parameter vector $\boldsymbol{\theta}$. To simplify notation we will write $\boldsymbol{\Delta}_{\mathbf{0}}$ and $\widehat{\boldsymbol{\Delta}}$ for $\boldsymbol{\Delta}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}_n)$, respectively. Likewise, we write $\boldsymbol{\sigma}_0$ and $\hat{\boldsymbol{\sigma}}_n$ for $\boldsymbol{\sigma}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}_n)$. Using elements of matrix calculus and asymptotic arguments as shown in appendix A, we get the following result on the asymptotic distribution of the MD estimator (Browne, 1984, Proposition 2):

Theorem 1. *Suppose $\hat{\boldsymbol{\theta}}_n$ is a MD estimator and that* (3) *holds. Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \xrightarrow{d} \quad N(\mathbf{0}, \boldsymbol{\Omega}).$$

If the estimator is correctly specified as given by (7) *then*

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}_{OPT} = (\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Delta}_0)^{-1}. \tag{8}$$

If the estimator is not correctly specified, then

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}_{SW} = (\boldsymbol{\Delta}_0' \boldsymbol{V} \boldsymbol{\Delta}_0)^{-1} \boldsymbol{\Delta}_0' \boldsymbol{V} \boldsymbol{\Gamma} \boldsymbol{V} \boldsymbol{\Delta}_0 (\boldsymbol{\Delta}_0' \boldsymbol{V} \boldsymbol{\Delta}_0)^{-1}.$$

So for any MD estimator $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is multivariate normal with a null mean vector and covariance matrix given by $\boldsymbol{\Omega}_{SW}$ above. The matrix $\boldsymbol{\Omega}_{SW}$ is commonly known as a sandwich- type covariance matrix, giving rise to robust "sandwich" standard errors. Standard error estimates of the MD estimator can now be obtained from the square roots of the diagonal elements of $\hat{\boldsymbol{\Omega}}$. When the estimator is correctly specified $\boldsymbol{\Omega}_{SW}$ reduces to $\boldsymbol{\Omega}_{OPT}$. In that case the estimator is optimal in the sense that it has the lowest possible variance among all MD estimators.

Example (continued)**.** *In our factor model $\boldsymbol{\theta}$ is simply the loading parameter λ and $\boldsymbol{\Delta}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\lambda) = (2\lambda, 2\lambda, 2\lambda)'$. Clearly it follows from the data-generating process* (4) *that $\lambda_0 = 1$ so $\boldsymbol{\Delta}_0 = (2, 2, 2)'$. The asymptotic covariance matrix in* (8) *is*

$$\boldsymbol{\Omega}_{OPT} = \left[(2 \quad 2 \quad 2) \begin{pmatrix} 8 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 8 \end{pmatrix}^{-1} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \right]^{-1} = [1.125]$$

and we have a univariate case of Theorem 1:

$$\sqrt{n}(\hat{\lambda}_n - 1) \quad \xrightarrow{d} \quad N(0, 1.125).$$

The asymptotic distribution of the residual vector

A key component in all goodness-of-fit chi-square statistics is the *residual vector*

$$\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n).$$

Intuitively the residual gives a measure of the goodness of fit of the model: If the model is good, then $\hat{\boldsymbol{\sigma}}_n$ should be quite close to $\boldsymbol{s}$, whereas for a less good model it would be further away.

Note that the residual is model-dependent while this is not the case for $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$, whose asymptotic distribution is solely a function of $\boldsymbol{\Gamma}$. The crucial insight - we relegate the mathematical details to appendix B - is that there is a close link between $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ and $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$:

$$\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) \stackrel{a}{=} \boldsymbol{P} \cdot \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0). \tag{9}$$

Here $\stackrel{a}{=}$ denotes "asymptotic equivalent to", which informally means that the left-hand and right-hand sides are virtually equal for large sample sizes. The matrix $\boldsymbol{P}$ is in general given by

$$\boldsymbol{P} \equiv \boldsymbol{I} - \boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V} \tag{10}$$

and it defines a linear transformation of a special kind, namely a *projection*. Hence the relation in (9) states that for large samples, the residual vector $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ is basically the result of projecting $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$ onto a subspace of lower dimension. Figure 2 gives a visual representation of the projection, where the residual vector is seen as the result of projecting $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$ onto a lower-dimensional subspace $\boldsymbol{X}$.

Figure 2 here.

It is interesting to note that the sampling distribution of the vector $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$ spans all directions in the space $\mathbb{R}^{p^*}$ in which it lives, while the residual vector for large samples tend to lie in a lower-dimensional subspace of $\mathbb{R}^{p^*}$. In statistical terms we say that the asymptotic distribution of $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ is *degenerate*. The dimension of the lower-dimensional space is equal to the *degrees of freedom of the model*, i.e. $p^* - q$. Informally one could say that for each free parameter in the model, the residual looses one degree of freedom. Define the *null space* of a matrix $\boldsymbol{A}$ as the set of vectors $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, and the *range* of a matrix $\boldsymbol{A}$ as the set of vectors $\boldsymbol{y}$ such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ for some vector $\boldsymbol{x}$. Then the full result is given in the following theorem.

Theorem 2. *Assume that the model holds. Then*

$$\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) \stackrel{a}{=} \boldsymbol{P} \cdot \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0), \tag{11}$$

where the $p^* \times p^*$ matrix $\boldsymbol{P}$ defined in (10) *represents an oblique projection onto the $p^* - q$-dimensional subspace*

$$\boldsymbol{X} = \mathit{Nullspace}(\boldsymbol{\Delta}_0'\boldsymbol{V})$$

along the subspace

$$\boldsymbol{Y} = \mathit{Range}(\boldsymbol{\Delta}_0).$$

Example (continued)**.** *In our example we have $p^* - q = 2$ so the residual resides asymptotically in a subspace of dimension 2, i.e. a plane. For correctly specified MD estimation we have*

$$\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^{-1} = (2 \quad 2 \quad 2) \begin{pmatrix} 8 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 8 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{9} & \frac{2}{9} & \frac{1}{9} \end{pmatrix}$$

and therefore $\boldsymbol{X}$ consists of all 3-vectors $\boldsymbol{x}$ such that $\begin{pmatrix} \frac{1}{9} & \frac{2}{9} & \frac{1}{9} \end{pmatrix} \cdot \boldsymbol{x} = 0$. In other words, $\boldsymbol{X}$ is the plane defined by

$$x_1 + 2x_2 + x_3 = 0.$$

In the next section we give a visual representation of this plane.

A remark on nested models. A model A is said to be *nested in the parameter sense* within a model B if the freely estimated parameters in model A is a subset of the freely estimated parameters in model B. Hence one can go from model B to model A by adding restrictions on some of the free parameters in B. For such nested models there exists an interesting relation between the subspaces $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$.

Proposition 1. *Suppose model A is parameter nested within model B, and that both models are correctly specified. Then* $\boldsymbol{X}_B \subset \boldsymbol{X}_A$.

The proof can be found in appendix C.

3 Visualization of simulated residuals

In this section we study simulations based on our simple one-factor model and a related model. The fact that these models include only $p = 2$ manifest variables and consequently that $p^* = 3$ allows us to visualise the residual vector $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ in 3-dimensional space. We focus on visualizing the residuals in relation to the subspace $\boldsymbol{X}$ as defined in Theorem 2. Our goal is to visually test how good an approximation equation (11) is across various models, estimation methods and sample sizes.

Three factors are incorporated into the design of the simulation study: model, estimation method and sample size. The model described in figure (1) will be referred to as Model 1, while a less restricted model will be referred to as Model 2.

Because ML and GLS are the most popular methods they were chosen as estimation methods. Sample sizes of 50, 250 and 1000 were investigated in the study. These sample sizes vary from a minimum requirement for SEM analysis through typical sample sizes for SEM and up to a large sample size. Simulation of random samples and estimation were done using the Lisrel/Prelis package (Joreskog, Sorbom, Du Toit, & Du Toit, 2000).

The random sample was generated according to the data-generating process described in the example on page 5. We remark that this ensures that our models are correctly specified and that the normality assumption

on the observables holds. For each sample size we generated 100 sample covariance matrices. The random samples were then used to fit our two models, and obtain the fitted residuals. This resulted in 100 residual 3-vectors for each model, estimation method and sample size. These 3-vectors were then imported into the Matlab package for visualization of the 100 residuals in a three-dimensional scatterplot. In the scatterplots we also plotted the subspace $\boldsymbol{X}$ referred to in Theorem 2.

In Figure 3 the 3D scatterplot is given for the residuals when the sample size of the 100 simulated datasets is $n = 250$. The GLS estimation was employed on Model 1. In Figure 3(a) one can see the plane $\boldsymbol{X}$ from an oblique angle. The residuals are scattered not far off the plane, as expected. However it is difficult to discern the precise location of the residuals, and an edge-on view as in Figure 3(b) offers a better picture of how the residuals are placed relative to the plane $\boldsymbol{X}$.

Figure 3 here.

Therefore the edge-on view is used in the following figures. However, to get a proper idea of the distribution of the residuals one should rotate the scatterplot. We provide rotation clips for the figures presented here at `http://home.bi.no/a0510192/wald`.

Model 1

As was seen in the example on page 9, for Model 1 $\boldsymbol{X} = \boldsymbol{X}_1$ is the plane defined by $x + 2y + z = 0$. In general, for no estimation method will equation (11) hold exactly for a finite sample size. However, in our particular case with Model 1 and maximum likelihood (ML) estimation, it is remarkable that the residuals fit tightly onto the plane for all sample sizes. This is shown in part F of the appendix.

For GLS estimation, however, as was seen in Figure 3, equation (11) does not hold for finite sample sizes. Figure 4 gives edge-on views of the GLS residuals for $n = 50$ and $n = 1000$. As expected, we see that for the larger sample size the residuals tend to lie closer to $\boldsymbol{X}_1$.

Figure 4 here.

Model 2

In this model the constraint $\lambda_1 = \lambda_2$ is removed from Model 1. Model 2 has $q = 2$ free parameters, namely λ_1 and λ_2, and hence $3 - 2 = 1$ degrees of freedom. Model 2 is depicted in figure (5).

Figure 5 here.

For Model 2 the subspace $\boldsymbol{X}_2 = \text{Nullspace}(\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1})$ is one-dimensional, i.e. $\boldsymbol{X}_2$ is a line through the origin. For Model 2 we have

$$\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix}' \begin{pmatrix} 8 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 8 \end{pmatrix}^{-1} = \frac{1}{9} \begin{pmatrix} 2 & 1 & -1 \\ -1 & 1 & 2 \end{pmatrix},$$

and hence $\boldsymbol{X}_2$ is the intersection of the two planes defined by $2x+y-z=0$ and $-x+y+2z=0$. In other words $\boldsymbol{X}_2$ is the line that passes through the origin trough the point $(1\,,-1,\,1)$. Note that this line $\boldsymbol{X}_2$ is contained in the plane $\boldsymbol{X}_1$ as predicted by the the discussion on page 9, since Model 1 is nested in Model 2.

Figure 6 here.

Figures 6(a) and 6(b) gives the ML residuals for $n=50$ and $n=250$, and we see that the residuals are closer to $\boldsymbol{X}_2$ for the larger sample size, as predicted by equation (11). In figures 6(c) and 6(d) we see the same pattern for the GLS residuals. Based on these samples it is not possible conclude which estimation method gives residuals closest to $\boldsymbol{X}_2$.

4 The residual-based test statistic

With Theorem 2 giving the asymptotic behavior of the residuals, we are now ready to study the residual-based test statistic. We first present a crucial proposition on the distribution of quadratic forms and then review the classical Wald test.

Wald's classical method for simple hypotheses

Let us first state a well-known property of quadratic forms. Let $\boldsymbol{y}=(y_1,\ldots,y_d)$ denote a random d-vector which is distributed according to the d-variate normal distribution, denoted by $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and the covariance matrix $\boldsymbol{\Sigma}$ is nonsingular. Since $\boldsymbol{\Sigma}^{-1}$ is positive definite there exists a matrix, denoted by $\boldsymbol{\Sigma}^{-\frac{1}{2}}$, such that $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}=\boldsymbol{\Sigma}^{-1}$.Now,

$$(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})=(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu})=\boldsymbol{z}'\boldsymbol{z},$$

where the standardized vector $\boldsymbol{z}=\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu})$ is normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}}=\boldsymbol{I}$. The right-hand side $\boldsymbol{z}'\boldsymbol{z}=\Sigma z_i^2$ is a sum of d independent squares of standard normal variables z_i. Such a sum of independent squared standard normal variables is per definition distributed as a chi-square with d degrees of freedom, denoted by $\chi^2(d)$, and we can state the following proposition:

Proposition 2. *Suppose that $\boldsymbol{y}$ is a d-vector which is distributed as $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is nonsingular. Then the quadratic form*

$$(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

is distributed as $\chi^2(d)$.

The residual-based goodness-of-fit statistic used in SEM is based on the same idea used by Wald (1943) for testing simple hypotheses. Wald's method in its simplest form is used to test whether a q-dimensional population parameter $\boldsymbol{\theta}$ is equal to some constant $\boldsymbol{\theta}_0$, i.e. to test the hypothesis

$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Let n denote the sample size and suppose $\hat{\boldsymbol{\theta}}_n$ is an estimator for $\boldsymbol{\theta}$. We have indexed the estimator by the sample size n since Wald's method only attains its desired properties for large samples, i.e. as $n \to \infty$. A crucial assumption is that the estimator is asymptotically normal. That is, the assumption is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \xrightarrow{d} \quad N(\mathbf{0}, \boldsymbol{\Omega}), \tag{12}$$

where $\boldsymbol{\Omega}$ is the nonsingular asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. To construct a measure of the discrepancy between $\boldsymbol{\theta}_0$ and the estimated $\hat{\boldsymbol{\theta}}_n$ let us start with a consistent estimator $\hat{\boldsymbol{\Omega}}_n$ of $\boldsymbol{\Omega}$. In many cases $\boldsymbol{\Omega}$ can be consistently estimated from the information matrix. The continuous mapping theorem in large-sample theory (e.g., Vaart, 2000, Theorem 2.3) states that if $\boldsymbol{z}_n$ is a sequence of random vectors that converges in distribution to $\boldsymbol{z}$, then for a continuous function g it holds that $g(\boldsymbol{z}_n)$ converges in distribution to $g(\boldsymbol{z})$. It then follows from Proposition 2 that

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \hat{\boldsymbol{\Omega}}_n^{-1} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \xrightarrow{d} \quad \chi^2(q).$$

Therefore, if H_0 holds, then $W_n = n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \hat{\boldsymbol{\Omega}}_n^{-1} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically χ^2-distributed with q degrees of freedom. Wald's method is simply to use the scalar W_n as a measure of discrepancy between the observed value $\hat{\boldsymbol{\theta}}_n$ and the proposed value $\boldsymbol{\theta}_0$. Values of W_n that exceed the critical value lead to the rejection of the null hypothesis. The critical value can be found, since the (asymptotic) distribution of W_n is known to be chi-square if the null hypothesis holds.

Wald's method can be used in conjunction with different estimation methods. ML estimation is a popular choice, but the only requirement for the method to be asymptotically valid is that the estimator approaches normality, i.e. that (12) holds. It has been noted that for small samples the estimator may be far from normally distributed. See Fears, Benichou, and Gail (1996) and Pawitan (2000) for situations where the Wald test exhibits poor power.

In structural equation modeling the null hypothesis is not of the simple form $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Rather H_0 states that the model is well-specified, meaning that there exists a parameter vector $\boldsymbol{\theta}_0$ such that the model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ equals the population covariance matrix $\boldsymbol{\Sigma}$. Hence Wald's original method is not directly suitable to test goodness of fit in SEM. In the following section we will show how Wald's idea of using an estimator that satisfies (12) to obtain a chi-square test statistic can be extended to construct a goodness of fit test for SEM.

Derivation of the residual-based test statistic

As described in Theorem 2, the residual is a p^*-vector that asymptotically lies in a subspace of lower dimension. Moreover, since the residual in (11) is a linear transformation of $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$, which is asymptotically normally distributed by assumption (3), the residual is also asymptotically normally

distributed. However, this normal distribution is *degenerate*, since the asymptotic covariance matrix of $\sqrt{n}(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)$ is singular; it has rank p^*-q. This means that the assumption of nonsingularity made in Proposition 2 is not met. However, this can be remedied by linearly mapping the residual onto the lower-dimensional space $\mathbb{R}^{p^*-q}$. This operation involves the notion of orthogonal complement.

Given an estimate $\hat{\boldsymbol{\theta}}_n$ the corresponding Jacobian $\widehat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}_n)$ is assumed to be of full column rank, namely q. An *orthogonal complement* of $\widehat{\boldsymbol{\Delta}}$ is a $p^* \times (p^*-q)$ matrix $\widehat{\boldsymbol{\Delta}}_c$ of full column rank such that $\widehat{\boldsymbol{\Delta}}_c'\widehat{\boldsymbol{\Delta}} = 0$. This basically means that any column of $\widehat{\boldsymbol{\Delta}}_c$ is orthogonal to any column of $\widehat{\boldsymbol{\Delta}}$. If we now multiply the residual by $\widehat{\boldsymbol{\Delta}}_c'$ on the left-hand side we get the following vector of dimension p^*-q:

$$\sqrt{n}\widehat{\boldsymbol{\Delta}}_c'(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n).$$

The main observation now is that, in contrast to $\sqrt{n}(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)$, this vector has asymptotically a non-degenerate normal distribution. This makes it possible to apply Proposition 2, since the asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\Delta}}_c(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)$ is non-singular, i.e. invertible. The algebraic details can be found in the appendix, part D. Consequently from Proposition 2 with $\sqrt{n}\widehat{\boldsymbol{\Delta}}_c'(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)$ taking the role of $\boldsymbol{y}$ we obtain the main result (Browne, 1984, Proposition 4):

Theorem 3. *Suppose our model is correct and that we estimate $\boldsymbol{\theta}_0$ by any (not necessarily correctly specified) MD estimator. Let $\widehat{\boldsymbol{\Gamma}}$ be a consistent estimator of $\boldsymbol{\Gamma}$. Then the residual-based statistic*

$$T_B = n(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)'\,\widehat{\boldsymbol{\Delta}}_c\,(\widehat{\boldsymbol{\Delta}}_c'\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Delta}}_c)^{-1}\,\widehat{\boldsymbol{\Delta}}_c'\,(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n) \tag{13}$$

is asymptotically distributed as the χ^2 distribution with p^*-q degrees of freedom.

We stress that the residual-based test statistic is asymptotically a chi-square regardless of the estimation method employed. Hence it is robust to non-normality even when used in conjunction with normal-theory based estimators like ML and GLS. Remark also that although the orthogonal complement matrix $\widehat{\boldsymbol{\Delta}}_c$ is not unique, the value of T_B in (13) does not depend on the choice of $\widehat{\boldsymbol{\Delta}}_c$. See appendix part D for details.

Finally, we now use Theorem (3) to deduce the asymptotic distribution of the MFF statistic

$$n\hat{F} = nF(\hat{\boldsymbol{\theta}}_n \mid \boldsymbol{V}_n)$$

where $\hat{\boldsymbol{\theta}}_n$ is the minimizer of F as given in (5). Note that main difference between the formulas in (13) and (5) is the presence of the orthogonal complement matrix. However, when the estimator $\hat{\boldsymbol{\theta}}_n$ is obtained by minimizing F this presence is redundant:

$$(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)'\boldsymbol{V}_n(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n) = (\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n)'\hat{\boldsymbol{\Delta}}_c(\hat{\boldsymbol{\Delta}}_c'\boldsymbol{V}_n^{-1}\hat{\boldsymbol{\Delta}}_c)^{-1}\hat{\boldsymbol{\Delta}}_c'(\boldsymbol{s}-\hat{\boldsymbol{\sigma}}_n). \tag{14}$$

A proof of (14) can be found in part E of the appendix. Therefore, Theorem (3) implies the following corollary:

Corollary 1. *Suppose the MD estimation of* (5) *is correctly specified. Then* $n\hat{F}$ *is asymptotically distributed as the* χ^2 *distribution with* $p^* - q$ *degrees of freedom.*

Consequently, with ML and GLS estimation the minimum fit function is asymptotically a chi-square for normal data. For non-normal data however, it can be shown that the minimum fit function is asymptotically a weighted sum of chi-squares. To partly remedy this departure from the reference chi-square distribution, Satorra and Bentler (1994) proposed a scaling of the minimum fit function that is asymptotically correct in mean for non-normal data.

Psychological data are often non-normal, in fact Micceri (1989) investigated 440 large-sample achievement and psychometric measures, and found all to be significantly nonnormal at the $\alpha = 0.01$ significance level. So there is definitely a need for test statistics that do not require the assumption of multivariate normality. The residual-based statistic T_B is a candidate for such a test statistic, but is relatively unknown. The reason is that the few simulation studies (e.g., Bentler & Yuan, 1998, 1999; Nevitt & Hancock, 2004) in SEM literature indicates that T_B performs poorly for small to moderate sample sizes. It tends to overreject true models. This issue has been studied in several articles by Bentler and Yuan (1999, 1998) which propose several corrections to T_B for small sample sizes.

5 Concluding remarks

In this paper we have studied the residual-based statistic T_B for goodness-of-fit in covariance structure analysis. T_B may be used as an asymptotically distribution free statistic with a theoretical elegance not found with other test statistics like $n\hat{F}$: it follows a known sampling distribution without assuming multivariate normality of the data. In fact, we have showed that T_B is asymptotically distributed as a chi-square with $p^* - q$ degrees of freedom. This holds in general for any MD estimation method, correctly specified or not. To explain why this holds we have focused on the residual vector $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ and demonstrated that it is asymptotically degenerate, i.e. although the residual is a p^*-vector it tends to reside in a $p^* - q$ dimensional subspace when the sample size increases. The treatment of the residuals is general in nature and helps understand the concept of degrees of freedom. We have proved that the residuals are constrained by MD estimation to asymptotically live in a subspace of dimensionality equal to the degrees of freedom of the model.

To exemplify the theory and to visualize the residual vectors we study a very simple model with only two observed variables. With two observed variables the residual vector resides in three-dimensional space and is readily available for visualization. In the simulation study the two observed variables follow a multivariate normal distribution. However, the theoretical results in this paper do not assume normality, in fact we only rely on very weak distributional assumptions.

Our discussion and results are based on the assumption that the model holds. This assumption simplifies the technical arguments, but it is often criticized for being unrealistic. In reality any model will at best approximate the process which underlies the generation of observed variables. To somehow ease the assumption of a well-specified model one could apply the device of a sequence of local alternatives to the null hypothesis, i.e. a sequence of population covariance matrices that converges to a population covariance matrix in which the model holds. This relaxed assumption of the correctness of the model is employed in Browne (1984), with analysis following largely the same lines as carried out in this exposition. The main conclusion is that T_B is then asymptotically distributed as a *non-central* chi-square with $p^* - q$ degrees of freedom.

Acknowledgement. *The authors wish to thank the associate editor and three anonymous referees for their helpful comments.*

Appendix

We make the mild assumption that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is bounded in probability. That is, we assume that for all $\epsilon > 0$ there exists a number M such that $P(\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \| > M) < \epsilon$ for all n.

A: Proof of Theorem 1

Let the gradient of $F(\boldsymbol{\theta} \mid \boldsymbol{V}_n)$ be denoted by

$$\dot{F}(\boldsymbol{\theta} \mid \boldsymbol{V}_n) \equiv \left(\frac{\partial F(\boldsymbol{\theta} \mid \boldsymbol{V}_n)}{\partial \theta_1}, \ldots, \frac{\partial F(\boldsymbol{\theta} \mid \boldsymbol{V}_n)}{\partial \theta_q} \right)',$$

We assume that $\boldsymbol{\Delta_0}$ and $\widehat{\boldsymbol{\Delta}}$ have full rank.

Applying the chain rule in matrix calculus (e.g., Magnus & Neudecker, 1999) the gradient can be expressed as

$$\dot{F}(\boldsymbol{\theta} \mid \boldsymbol{V}_n) = -2\boldsymbol{\Delta}(\boldsymbol{\theta})' \boldsymbol{V}_n \left(\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}) \right). \tag{15}$$

Since $\hat{\boldsymbol{\theta}}_n$ is the MD estimator we have $\dot{F}(\hat{\boldsymbol{\theta}}_n \mid \boldsymbol{V}_n) = 0$ and

$$0 = \widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n \left(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n \right) = \widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n \left(\boldsymbol{s} - \boldsymbol{\sigma}_0 - (\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}_0) \right)$$

which we rewrite as

$$\widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n (\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}_0) = \widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n (\boldsymbol{s} - \boldsymbol{\sigma}_0). \tag{16}$$

On the left-hand side, Taylor expansion of $\boldsymbol{\sigma}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ gives

$$\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}_0 = \boldsymbol{\Delta}_0 (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{17}$$

where the remainder function r satisfies $\lim_{\boldsymbol{u} \to 0} r(\boldsymbol{u}) / \|\boldsymbol{u}\| = 0$ for a q-vector $\boldsymbol{u}$ and the Euclidean norm $\|\cdot\|$ (see Magnus and Neudecker (1999) for multivariate Taylor expansion). After multiplying (16) with $\sqrt{n}$ and combining with (17) we get

$$\widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n (\boldsymbol{\Delta}_0 \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \sqrt{n} r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)) = \widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0). \tag{18}$$

For the last term on the left-hand side it holds that

$$\sqrt{n} r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \frac{r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|} \quad \xrightarrow{P} \quad 0$$

since $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is bounded in probability. Therefore

$$\widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n \boldsymbol{\Delta}_0 \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{a}{=} \widehat{\boldsymbol{\Delta}}' \boldsymbol{V}_n \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0), \tag{19}$$

where $\stackrel{a}{=}$ stands for "asymptotically equivalent", meaning that the difference between the left- and right hand sides converges in probability towards zero. Let us assume that the estimator is correctly specified, i.e. that $\boldsymbol{V} = \boldsymbol{\Gamma}^{-1}$. Since $\widehat{\boldsymbol{\Delta}} \xrightarrow{P} \boldsymbol{\Delta}_0$ we can replace $\widehat{\boldsymbol{\Delta}}$ by $\boldsymbol{\Delta}_0$ in (19) and left-multiply by $(\boldsymbol{\Delta}_0' \boldsymbol{V}_n \boldsymbol{\Delta}_0)^{-1}$ to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{a}{=} (\boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Delta}_0)^{-1} \boldsymbol{\Delta}_0' \boldsymbol{\Gamma}^{-1} \sqrt{n} \left(\boldsymbol{s} - \boldsymbol{\sigma}_0 \right). \tag{20}$$

Now if $\boldsymbol{x}$ is a random vector with covariance matrix $\boldsymbol{C}$, then $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$ has the covariance matrix $\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}'$. So it follows from (20) that the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in probability towards

$$\boldsymbol{\Omega}_{OPT} = (\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1}\cdot\boldsymbol{\Gamma}\cdot\big((\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}'\boldsymbol{\Gamma}^{-1}\big)' = \big(\boldsymbol{\Delta}_0'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}_0\big)^{-1}$$

for correctly specified MD estimation. The result for $\boldsymbol{\Omega}_{SW}$ is obtained by replacing $\boldsymbol{\Gamma}^{-1}$ in (20) by $\boldsymbol{V}$.

B: Proof of Theorem 2

We assume that the model holds and focus on the asymptotic distribution of

$$\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) = \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0) - \sqrt{n}(\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}_0). \tag{21}$$

Using (17) again, together with the succeeding argument about the disappearance of the remainder gives us the asymptotic equivalence

$$\sqrt{n}(\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}_0) \stackrel{a}{=} \Delta_0\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

and combining this with equations (21) and (20), where we replace $\boldsymbol{\Gamma}^{-1}$ in (20) by $\boldsymbol{V}$, it follows that

$$\begin{aligned}\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) &\stackrel{a}{=} (\boldsymbol{I} - \boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V}\cdot\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)\\ &= \boldsymbol{P}\cdot\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0)\end{aligned}$$

where

$$\boldsymbol{P} \equiv \boldsymbol{I} - \boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V} \tag{22}$$

is a projection matrix. This follow from the fact that $\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P}^2 = \boldsymbol{P}$, which can be shown by straighforward calculation.

The range $\boldsymbol{X}$ of $\boldsymbol{P}$ consists of exactly those vectors $\boldsymbol{x}$ such that $\boldsymbol{P}\boldsymbol{x} = \boldsymbol{x}$:

$$\begin{aligned}&(\boldsymbol{I} - \boldsymbol{\Delta}_0\left(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0\right)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V})\boldsymbol{x} = \boldsymbol{x}\\ \iff\quad &\boldsymbol{\Delta}_0\left(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0\right)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{x} = 0\\ \iff\quad &\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{x} = 0,\end{aligned}$$

where we have used that $\boldsymbol{V}$ is nonsingular and that $\boldsymbol{\Delta}_0$ has full column-rank q. Hence, $\boldsymbol{P}$ is the projection onto the subspace $\boldsymbol{X} = \text{Nullspace}(\boldsymbol{\Delta}_0'\boldsymbol{V})$ along the subspace

$$\boldsymbol{Y} = \text{Range}(\boldsymbol{\Delta}_0\left(\boldsymbol{\Delta}_0'\boldsymbol{V}\boldsymbol{\Delta}_0\right)^{-1}\boldsymbol{\Delta}_0'\boldsymbol{V}) = \text{Range}(\boldsymbol{\Delta}_0).$$

The last identity again follows from the fact that $\boldsymbol{\Delta}_0$ has full rank. The dimension of $\boldsymbol{X}$ is $p^* - q$ since $\boldsymbol{\Delta}_0'\boldsymbol{V}$ represents a linear transformation from $\mathbb{R}^{p^*}$ onto $\mathbb{R}^q$ of rank q. The theorem follows.

As a final note, it is not suprising that the residual asymptotically resides in the subspace $\boldsymbol{X}$ in light of equation (15). That equation states that the residual vector is in the nullspace of the matrix $\widehat{\boldsymbol{\Delta}}_n'\boldsymbol{V}_n$, which converges in probability to $\boldsymbol{\Delta}_0'\boldsymbol{V}$.

C: Nested models

Suppose Model B has the free parameters $\theta_1, \ldots, \theta_q, \gamma_1, \ldots \gamma_r$. In Model A this parameter set must satisfy r equality constraints, and we assume that each equality constraint can be written as $\gamma_j = c_j(\theta_1, \ldots, \theta_q)$ for $j = 1, \ldots, r$ where the c_j are continuously differentiable functions. For instance, $c_j(\theta_1, \ldots, \theta_q) = \theta_i$ means that the free parameter γ_j in Model B is in Model A restricted to be equal to the parameter θ_i. The restrictions that makes Model A nested within Model B is therefore represented by the differentiable mapping $\boldsymbol{c}$ from $\mathbb{R}^q$ into $\mathbb{R}^{q+r}$ defined by:

$$(\theta_1, \ldots, \theta_q) \mapsto (\theta_1, \ldots, \theta_q, c_1(\theta_1, \ldots, \theta_q), \ldots, c_r(\theta_1, \ldots, \theta_q)).$$

Now let $\theta_1^0, \ldots, \theta_q^0, \gamma_1^0, \ldots \gamma_r^0$ be the unique parameter values such that $\sigma_B(\theta_1^0, \ldots, \theta_q^0, \gamma_1^0, \ldots \gamma_r^0) = \boldsymbol{\sigma}_0$. Since Model A is correctly specified $c_j(\theta_1^0, \ldots, \theta_q^0) = \gamma_j^0$ for $j = 1, \ldots, r$. Note that for Model A σ_A is the composite function of the Model B σ_B and the function $\boldsymbol{c}$:

$$\sigma_A(\theta_1, \ldots, \theta_q) = \sigma_B(\boldsymbol{c}(\theta_1, \ldots, \theta_q)).$$

Hence, the multivariable chain rule can be applied (see p.91 in (Magnus & Neudecker, 1999)):

$$\boldsymbol{\Delta}_{0A} = \boldsymbol{\Delta}_{0B} \cdot \boldsymbol{D}c(\theta_1^0, \ldots, \theta_q^0), \tag{23}$$

where $\boldsymbol{D}c$ is the differential of $\boldsymbol{c}$. Now, if $\boldsymbol{z} \in \boldsymbol{X}_B$, then $\boldsymbol{\Delta}'_{0B}\boldsymbol{\Gamma}^{-1}\boldsymbol{z} = \boldsymbol{0}$, so $\boldsymbol{\Gamma}^{-1}\boldsymbol{z}$ is orthogonal to the column space of $\boldsymbol{\Delta}_{0B}$. By (23) this column space contains the column space of $\boldsymbol{\Delta}_{0A}$ and hence $\boldsymbol{\Gamma}^{-1}\boldsymbol{z}$ is orthogonal to the column space of $\boldsymbol{\Delta}_{0A}$. Therefore $\boldsymbol{\Delta}'_{0A}\boldsymbol{\Gamma}^{-1}\boldsymbol{z} = \boldsymbol{0}$ and $\boldsymbol{z} \in \boldsymbol{X}_A$, and (2) follows.

D: Proof of Theorem 3

From (11) and $\widehat{\boldsymbol{\Delta}}_c \xrightarrow{P} \boldsymbol{\Delta}_{0c}$ we get

$$\begin{aligned}
\sqrt{n}\widehat{\boldsymbol{\Delta}}'_c(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) &\overset{a}{=} \widehat{\boldsymbol{\Delta}}'_c \left(\boldsymbol{I} - \boldsymbol{\Delta}_0 \left(\boldsymbol{\Delta}'_0 \boldsymbol{V} \boldsymbol{\Delta}_0\right)^{-1} \boldsymbol{\Delta}'_0 \boldsymbol{V}\right) \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0) \\
&\overset{a}{=} \boldsymbol{\Delta}'_{0c} \left(\boldsymbol{I} - \boldsymbol{\Delta}_0 \left(\boldsymbol{\Delta}'_0 \boldsymbol{V} \boldsymbol{\Delta}_0\right)^{-1} \boldsymbol{\Delta}'_0 \boldsymbol{V}\right) \sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}_0) \\
&= \sqrt{n}\boldsymbol{\Delta}'_{0c}(\boldsymbol{s} - \boldsymbol{\sigma}_0).
\end{aligned} \tag{24}$$

By assumption (3) $\sqrt{n}\boldsymbol{\Delta}'_{0c}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$ is asymptotically normally distributed, and its asymptotic covariance matrix is $\boldsymbol{\Delta}'_{0c}\boldsymbol{\Gamma}\boldsymbol{\Delta}_{0c}$. It therefore follows from (24) that $\sqrt{n}\widehat{\boldsymbol{\Delta}}'_c(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ is also asymptotically normally distributed:

$$\sqrt{n}\widehat{\boldsymbol{\Delta}}'_c(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) \quad \xrightarrow{d} \quad N(0, \boldsymbol{\Delta}'_{0c}\boldsymbol{\Gamma}\boldsymbol{\Delta}_{0c}). \tag{25}$$

Now Proposition 2 is applicable, since the asymptotic covariance matrix $\boldsymbol{\Delta}'_{0c}\boldsymbol{\Gamma}\boldsymbol{\Delta}_{0c}$ is non-singular, i.e. invertible. This non-singularity stems from the fact that $\boldsymbol{\Delta}_{0c}$ has full rank, and that $\boldsymbol{\Gamma}$ is positive definite. Consequently from Proposition 1 with $\sqrt{n}\widehat{\boldsymbol{\Delta}}_c(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ taking the role of $\boldsymbol{y}$ we obtain our main result.

To see that the choice of orthogonal complement does not change the value of T_B, note that any two orthogonal complement matrices $\boldsymbol{\Delta}_c^1$ and $\boldsymbol{\Delta}_c^2$ can be related by a non-singular p^*-q by p^*-q matrix $\boldsymbol{Q}$: $\boldsymbol{\Delta}_c^1 = \boldsymbol{\Delta}_c^2 \cdot \boldsymbol{Q}$. Using this relation and with the aid of basic matrix algebra it follows that

$$\boldsymbol{\Delta}_c^1 (\boldsymbol{\Delta}_c^{1\prime} \boldsymbol{\Gamma} \boldsymbol{\Delta}_c^1)^{-1} \boldsymbol{\Delta}_c^{1\prime} = \boldsymbol{\Delta}_c^2 (\boldsymbol{\Delta}_c^{2\prime} \boldsymbol{\Gamma} \boldsymbol{\Delta}_c^2)^{-1} \boldsymbol{\Delta}_c^{2\prime}.$$

E: Proof of equation (14)

Without loss of generality we assume that the column vectors in the orthogonal complement has been normalized: $\hat{\boldsymbol{\Delta}}_c' \cdot \hat{\boldsymbol{\Delta}}_c = \boldsymbol{I}$. If the estimator $\hat{\boldsymbol{\theta}}_n$ is obtained by minimizing F, then the gradient in (15) must be zero: $\widehat{\boldsymbol{\Delta}}' \cdot \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) = 0$. This implies the existence of a $p^* - q$ vector $\boldsymbol{u}$ such that $\boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) = \hat{\boldsymbol{\Delta}}_c \cdot \boldsymbol{u}$. Next observe that

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \hat{\boldsymbol{\Delta}}_c \cdot \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) &= \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \hat{\boldsymbol{\Delta}}_c \hat{\boldsymbol{\Delta}}_c' \hat{\boldsymbol{\Delta}}_c \cdot \boldsymbol{u} \\
&= \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \hat{\boldsymbol{\Delta}}_c \boldsymbol{u} = \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) \\
&= \hat{\boldsymbol{\Delta}}_c' (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n).
\end{aligned}$$

It follows that $\left(\hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \hat{\boldsymbol{\Delta}}_c \right)^{-1} \cdot \hat{\boldsymbol{\Delta}}_c' (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) = \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$, and we get

$$\begin{aligned}
(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)' \hat{\boldsymbol{\Delta}}_c \cdot (\hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n^{-1} \hat{\boldsymbol{\Delta}}_c)^{-1} \hat{\boldsymbol{\Delta}}_c' (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) &= (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)' \hat{\boldsymbol{\Delta}}_c \cdot \hat{\boldsymbol{\Delta}}_c' \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n) \\
&= (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)' \hat{\boldsymbol{\Delta}}_c \hat{\boldsymbol{\Delta}}_c' \hat{\boldsymbol{\Delta}}_c \boldsymbol{u} \\
&= (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)' \hat{\boldsymbol{\Delta}}_c \boldsymbol{u} \\
&= (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)' \boldsymbol{V}_n (\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n).
\end{aligned}$$

F: ML residuals for Model 1

Suppose the sample covariance matrix is $\boldsymbol{S} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, and consider the well-known maximum-likelihood fit function:

$$F_{ML} = \ln |\boldsymbol{\Sigma}(\lambda)| + tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\lambda)) + C.$$

Differentiating with respect to λ gives

$$\frac{dF_{ML}}{d\lambda} = \frac{8\lambda^3 + (4 - 2a - 2c - 4b)\lambda}{(2\lambda^2 + 1)^2}.$$

It follows that the ML estimate $\widehat{\lambda}$ satisfies

$$\widehat{\lambda}^2 = \frac{a + 2b + c - 2}{4}$$

and replacing this in the residual gives

$$\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n = \begin{pmatrix} a \\ b \\ c \end{pmatrix} - \begin{pmatrix} \widehat{\lambda}^2 + 1 \\ \widehat{\lambda}^2 \\ \widehat{\lambda}^2 + 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 3a - 2b - c - 2 \\ -a + 2b - c + 2 \\ -a - 2b + 3c - 2 \end{pmatrix}.$$

Clearly this vector satisfies $x + 2y + z = 0$, proving that the ML residual $\sqrt{n}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}_n)$ lies in the plane for any sample size n.

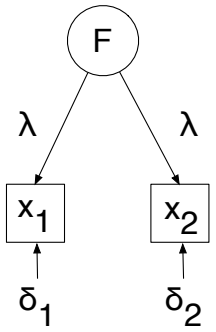


Figure 1: A simple factor model.

ULS	$2^{-1}\boldsymbol{K}_p^-(\boldsymbol{I}\otimes\boldsymbol{I})\boldsymbol{K}_p^{-\prime}$
GLS	$2^{-1}\boldsymbol{K}_p^-(\boldsymbol{S}^{-1}\otimes\boldsymbol{S}^{-1})\boldsymbol{K}_p^{-\prime}$
ML	$2^{-1}\boldsymbol{K}_p^-(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1})\boldsymbol{K}_p^{-\prime}$
ADF	$\widehat{\boldsymbol{A}}^{-1}$

Table 1: $\boldsymbol{V}_n$ for four estimators

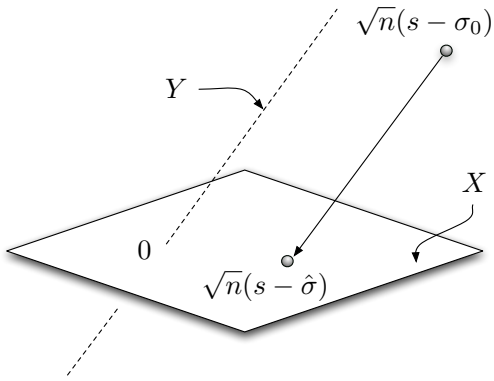
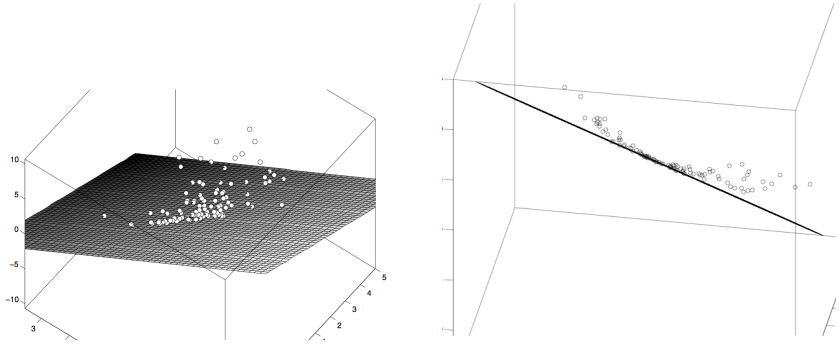


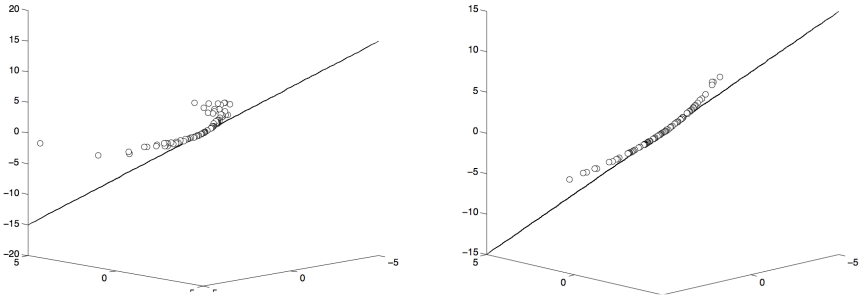
Figure 2: $\sqrt{n}(s-\hat{\boldsymbol{\sigma}}_n)$ is the projection of $\sqrt{n}(s-\boldsymbol{\sigma}_0)$.



(a) Oblique perspective.

(b) Edge-on perspective.

Figure 3: Model 1: GLS residuals for $n = 250$.



(a) $n = 50$. (b) $n = 1000$.

Figure 4: Model 1: GLS residuals.

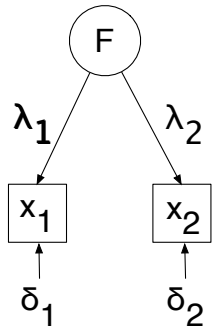
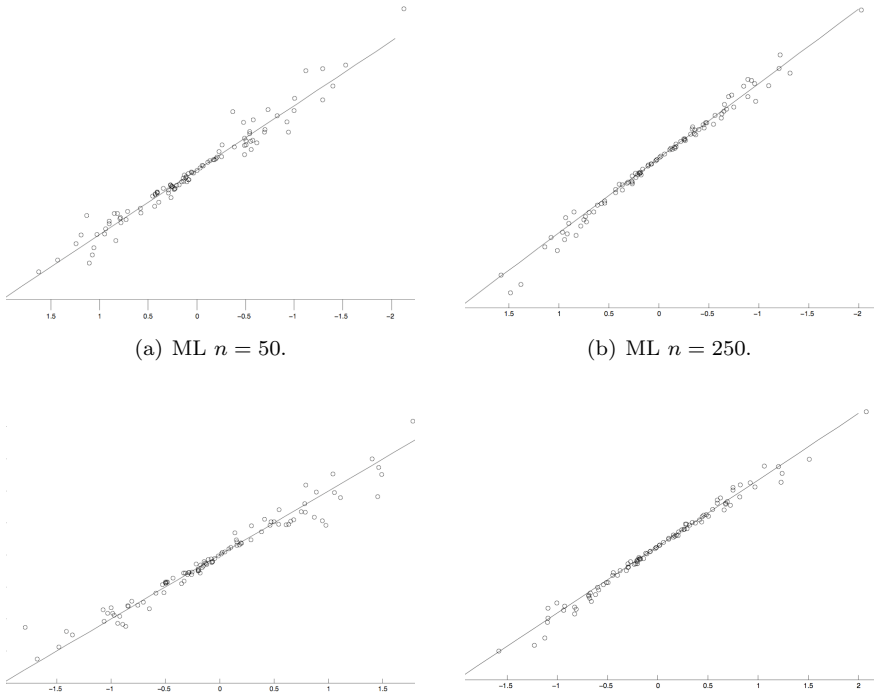


Figure 5: Model 2.



(a) ML $n = 50$.

(b) ML $n = 250$.

(c) GLS $n = 50$.

(d) GLS $n = 250$.

Figure 6: Residuals for Model 2.

References

Anderson, T. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley-Interscience.

Bentler, P., & Yuan, K. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197.

Bentler, P., & Yuan, K.-H. (1998). Normal theory based test statistics in structural equation modeling. *British journal of mathematical and statistical psychology*, *51*, 289-309.

Browne, M. (1974). Generalized least-squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*, 1-24.

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British journal of mathematical and statistical psychology*, *37*, 62-83.

Cai, L., & Lee, T. (2009). Covariance structure model fit testing under missing data: An application of the supplemented em algorithm. *Multivariate Behavioral Research*, *44*(2), 281–304.

Fears, T., Benichou, J., & Gail, M. (1996). A reminder of the fallibility of the wald statistic. *The American Statistician*, *50*(3).

Joreskog, K., Sorbom, D., Du Toit, S., & Du Toit, M. (2000). *Lisrel 8: New statistical features.* Chicago, IL: Scientific Software International.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166.

Nevitt, J., & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478.

Pawitan, Y. (2000). A reminder of the fallibility of the wald statistic: Likelihood explanation. *American Statistician*, *54*(1), 54–56.

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

Satorra, A. (2003). Power of chi-square goodness-of-fit tests in structural equation models: the case of non-normal data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. Meulman (Eds.), *New developments in psychometrics* (p. 57-68). Tokyo: Springer Verlag.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. V. Eye &

C. Clogg (Eds.), *Latent variable analysis: applications for developmental research* (p. 399-419). Newbury Park, CA: Sage.

Savalei, V., & Bentler, P. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 477–497.

Shapiro, A. (2007). Statistical inference of moment structures. In *Handbook of computing and statistics with applications* (Vol. 1, p. 229-260). Elsevier B.V.

Vaart, A. V. der. (2000). *Asymptotic statistics*. Cambridge University Press.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482.

Yuan, K., & Bentler, P. (2007). Structural equation modeling. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26: Psychometrics, p. 297-348). North-Holland.

Yuan, K., Bentler, P., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis. *Sociological Methods & Research*, *34*(2), 240-258.