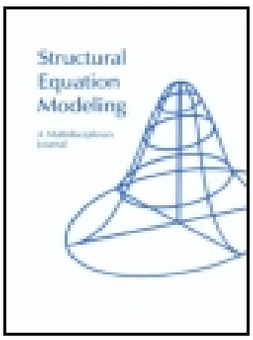

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Testing Mean and Covariance Structures with Reweighted Least Squares

Bang Quan Zheng & Peter M. Bentler

To cite this article: Bang Quan Zheng & Peter M. Bentler (2021): Testing Mean and Covariance Structures with Reweighted Least Squares, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2021.1977649

To link to this article: https://doi.org/10.1080/10705511.2021.1977649

Published online: 15 Oct 2021.

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2021.1977649

Testing Mean and Covariance Structures with Reweighted Least Squares

Bang Quan Zheng and Peter M. Bentler

University of California, Los Angeles

ABSTRACT

Chi-square tests based on maximum likelihood (ML) estimation of covariance structures often incorrectly over-reject the null hypothesis: $\Sigma = \Sigma(\theta)$ when the sample size is small. Reweighted least squares (RLS) avoids this problem. In some models, the vector of parameter must contain means, variances, and covariances, yet whether RLS also works in mean and covariance structures remains unexamined. This research extends RLS to mean and covariance structures, evaluating a generalized least squares function with ML parameter estimates. A Monte Carlo simulation study was carried out to examine the statistical performance of ML vs RLS with multivariate normal data. Based on empirical rejection frequencies and empirical averages of test statistics, this study shows that RLS performs much better than ML in mean and covariance structure models when sample sizes are small, whereas it does not perform better than ML to reject misspecified models.

KEYWORDS

Mean and covariance structure; reweighted least squares; goodness-of-fit test; structural equation models; small sample size

Introduction

Structural equation modeling (SEM) statistics such as those from maximum likelihood (ML) and generalized least squares (GLS) are based on asymptotic properties, in which sample sizes are assumed to be very large. Then the associated conventional goodness-of-fit test for model adequacy asymptotically follows a standard χ^2 distribution. This property holds for covariance structures and for joint mean and covariance structures. Unfortunately, in actual applications in social science research, particularly in longitudinal data with growth curve modeling (GCM), violation of asymptotic sample sizes is typical. As a result, the most widely utilized ML χ^2 goodness-of-fit test (Jöreskog, 1969) too often incorrectly rejects the null hypothesis even when the model specification is correct (e.g., Arruda & Bentler, 2017; Hayakawa, 2019; Jalal & Bentler, 2018). Additional contributors to model over-rejection include the number of variables, so that when the size of the covariance matrix is large, the correct null hypothesis is excessively rejected (Moshagen, 2012; Shi et al., 2018), and when the number of free parameters or degrees of freedom of the model are large, model over-rejection occurs (Herzog et al., 2007; Hoogland & Boomsma, 1998; Jackson, 2003). Finally, violation of multivariate normality when using normal-theory-based tests such as ML also results in excessive model rejection (e.g., Hu et al., 1992; Yuan & Bentler, 1997). This paper limits its scope to the effects of sample size on GLS, ML, and RLS test statistics in correctly and misspecified mean and covariance structure models with normal data.

Building on the reweighted least squares (RLS) approach introduced by Browne (1974) for covariance structures, and reintroduced by Hayakawa (2019), this research extends RLS to mean and covariance structures and studies its performance as compared to GLS and ML for its null hypothesis performance and its power to reject misspecified models.

The method undertaken in this research is quite straightforward. It relies on Monte Carlo Simulation to draw different sample sizes from N = 50 to 10,000 to compare the performance of chi-square model fit statistics from estimators of interest for both covariance structures as well as mean and covariance structures. Using 1,000 replications at each sample size, we find that RLS outperforms GLS and ML on mean and covariance structures and offers highly consistent goodness-of-fit chi-square model tests across different sample sizes. In contrast, RLS does not perform better than ML to reject misspecified models.

This paper is organized as follows. It first reviews covariance structure analysis with its ML, GLS, and RLS test statistics. The next section reviews mean and covariance structures and develops RLS in this context. The subsequent section discusses data generation and the simulation procedures, followed by evaluation criteria and then results, including power analysis of these methods. The last section provides a discussion and conclusion.

Covariance structures

In this section, we review parameter estimates and model fit tests with covariance structures. Let $\{x_1, \ldots, x_N\}$ be a random sample of x, with the x_i identically and independently distributed according to a multivariate normal distribution $\boldsymbol{N}[\boldsymbol{0}, \boldsymbol{\Sigma}]$. We assume that $\boldsymbol{\Sigma}$ $(p \times p)$ is a matrix function of an unknown vector of population parameters $\boldsymbol{\theta}(q \times 1)$, with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. The unstructured sample covariance matrix is

CONTACT Bang Quan Zheng bangquan@ucla.edu Department of Psychology, Florida International University, Miami, FL 33199.
The work was conceptualized and implemented at UCLA but completed at Florida International University.

$$\boldsymbol{S} = \frac{1}{N-1}\Sigma_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})' \tag{1}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}(x_1, \ldots, x_N)$ is the sample mean. When the sample size N is large, the difference between $\frac{1}{N}$ and $\frac{1}{N-1}$ can be neglected. According to the Multivariate Central Limit Theorem (Anderson, 1984), the unbiased sample covariance matrix $\boldsymbol{S}$ is positive definite with probability 1 and converges to $\boldsymbol{\Sigma}$ in probability. The asymptotic distribution of $\boldsymbol{s} = \text{vech}(\boldsymbol{S})$ is

$$N^{1/2}(\boldsymbol{s} - \sigma(\boldsymbol{\theta})) \xrightarrow{L} n[0,\ 2\boldsymbol{K}_p'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{K}_p] \tag{2}$$

where "$\xrightarrow{L}$" denotes convergence in distribution, $\text{vech}(\boldsymbol{S})$ $\frac{p(p+1)}{2} \times 1$ may be expressed in terms of the $p^2 \times 1\,\text{vech}(\boldsymbol{S})$, and similarly for $\sigma(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, with $\boldsymbol{K}_p$ of order $p^2 \times p(p+1)/2$.

The specific covariance structure to be studied herein is the confirmatory factor analysis (CFA) model in deviation score form

$$\boldsymbol{x_i} = \boldsymbol{\Lambda}\boldsymbol{\xi_i} + \varepsilon_i, \qquad i = 1, \ldots, N$$

where $\boldsymbol{x}_i$ is a random sample, $\boldsymbol{\Lambda}$ $(p \times m)$ is a matrix of factor loadings, $\boldsymbol{\xi_i}$ $(p \times 1)$ is a vector of latent common factors, and $\varepsilon_i(p \times 1)$ is a vector of unique factors. With the usual CFA assumptions, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Phi}$ is the $m \times m$ covariance matrix of the common factors and $\boldsymbol{\Psi}$ is the $p \times p$ diagonal covariance matrix of unique factors. The unknown parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are elements of $\boldsymbol{\theta}$.

To estimate the unknown parameters in $\boldsymbol{\theta}$, we minimize an objective function $F[\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{S}]$ that measures the discrepancy between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{S}$. Functions relevant to this paper are ML (Jöreskog, 1969) and GLS (Browne, 1974). The ML function to be minimized is

$$F_{ML}(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\boldsymbol{S}| + tr\left(\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right) - p \tag{3}$$

leading to optimal parameter estimates

$$\widehat{\boldsymbol{\theta}}_{ML} = argmin\ F_{ML}(\boldsymbol{\theta}). \tag{4}$$

The associated goodness-of-fit test statistic is

$$T_{ML} = (N-1)F_{ML}\left(\widehat{\boldsymbol{\theta}}\right), \tag{5}$$

which asymptotically follows a chi-square distribution with degrees of freedom $df = p^* - q$, where $p^* = p(p+1)/2$ and q is the number of free parameters.

The GLS function was proposed by Browne (1974)

$$\begin{aligned} F_{GLS} &= 2^{-1}(\boldsymbol{s} - \sigma(\boldsymbol{\theta}))'(\boldsymbol{V} \otimes \boldsymbol{V})vec(\boldsymbol{s} - \sigma(\boldsymbol{\theta})) \\ &= 2^{-1}tr\left[\{(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))\boldsymbol{V}\}^2\right] \end{aligned} \tag{6}$$

where $\otimes$ is a Kronecker product and $\boldsymbol{V}$ is a constant or stochastic matrix that converges to a consistent positive definite estimator of $\boldsymbol{\Sigma}^{-1}$ (Lee, 2007). Typically in GLS, $V = S^{-1}$, and at the minimum of $\hat{F}_{GLS}$, one obtains parameter estimates $\widehat{\boldsymbol{\theta}}_{GLS}$ and the GLS test statistic $T_{GLS} = (N-1)F_{GLS}\left(\widehat{\boldsymbol{\theta}}\right)$ with $p^* - q\,df$.

Reweighted least squares

The RLS function (Browne, 1974, Prop. 7) is a special case of (6). The parameter estimates are taken as $\widehat{\boldsymbol{\theta}}_{ML}$, with $\widehat{\boldsymbol{\Sigma}}_{ML}$ used in (6) to yield the test statistic

$$T_{RLS} = \frac{N}{2}tr\left\{\left(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}}_{ML}\right)\widehat{\boldsymbol{\Sigma}}_{ML}^{-1}\right\}^2. \tag{7}$$

T_{RLS} asymptotically follows a chi-square distribution, that is, $T_{RLS} \xrightarrow{L} \chi^2_{df}$ as $N \to \infty$. The relationship between T_{ML} and T_{RLS} was shown by Browne (1974) to be

$$T_{ML} = T_{RLS} + B$$

$$B = N\Sigma_{k=3}^{\infty}\frac{1}{k}tr\left\{I_p - \boldsymbol{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right\}^k. \tag{8}$$

While the term B vanishes asymptotically, Hayakawa (2019) points out that although B can be positive or negative, B is mostly positive with large p. Also, as p increases, the relative magnitude of B to degrees of freedom also increases. Then if the test statistic T_{RLS} is close to its expected value, T_{ML} will tend to be too large. When sample size is sufficiently large, B will vanish, and the RLS and ML tests become equivalent. Equation (8) explains simulation results showing T_{RLS} can remain highly consistent across sample sizes while T_{ML} is too large in small samples (e.g., Hayakawa, 2019).

Mean and covariance structures

In their discussion of latent curve or GCM models, Bollen and Curran (2005) note that the methodology involves a simultaneous null hypothesis for means and covariances as functions of the more basic parameters $\boldsymbol{\theta}$

$$H_o : \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) \text{ and } \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}).$$

Specifically, we again consider the CFA model $\boldsymbol{x_i} = \boldsymbol{\Lambda}\boldsymbol{\xi_i} + \varepsilon_i$, but now to its covariance structure, we add the expectations $E(\boldsymbol{x_i}) = \boldsymbol{\mu}$, $E(\xi_i) = \boldsymbol{\mu}_\xi$, and $E(\varepsilon_i) = \boldsymbol{0}$. This results in the mean structure (MS)

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi, \tag{9}$$

implying that observed variable means are a linear combination of latent factor means with weights given by $\boldsymbol{\Lambda}$. While this structure could be evaluated against sample data using a type of GLS function (e.g., Yuan et al., 2019)

$$\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right),$$

this would ignore the simultaneous null hypothesis $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Thus, for ML estimation, we use the compound covariance and MS discrepancy function

$$T_{ML\_MS} = T_{ML} + (N-1)\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right)'\boldsymbol{\Sigma}_{ML}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right), \tag{10}$$

where T_{ML} was given in (5) and now $\boldsymbol{\theta}$ also contains the unknown factor mean parameters $\boldsymbol{\mu}_\xi$. At the minimum of (10), we obtain $\widehat{\boldsymbol{\theta}}_{ML}$ and the test statistics T_{ML_MS}, which is referred to χ^2_{df}, where $df = p^* - q$, where now $p^* = p + p(p+1)/2$ and the number of free parameters q now also contains the number of unknown factor means. The comparable covariance-mean structure GLS function to be minimized is

$$\boldsymbol{T_{GLS\_MS}} = \boldsymbol{T_{GLS}} + (\boldsymbol{N}-1)\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right)'\boldsymbol{S}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{\Lambda}\boldsymbol{\mu}_\xi\right), \tag{11}$$

where T_{GLS} was defined in association with (6). As was (10), at the minimum of (11), T_{GLS_MS} is referred to χ^2_{df}.

RLS extended to mean and covariance structures is parallel to (11) but involves no function minimization. Rather, the estimates $\widehat{\boldsymbol{\theta}}_{ML}$ obtained by minimizing (10) now include ML estimates $\widehat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{\xi}}$ that are used along with $\widehat{\boldsymbol{\Sigma}}_{ML}$ and $\hat{\boldsymbol{\Sigma}}^{-1}_{ML}$ in the combined covariance/mean test statistic

$$T_{RLS_{MS}} = T_{RLS} + (N-1)\left(\bar{\boldsymbol{x}} - \widehat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\xi}}\right)' \widehat{\boldsymbol{\Sigma}}^{-1}_{ML}\left(\bar{\boldsymbol{x}} - \widehat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\xi}}\right), \quad (12)$$

where T_{RLS} was previously given in (7).

Data generation and simulation

The data generation scheme for the simulation follows the structured means CFA model described previously, where $\boldsymbol{x_i} = \boldsymbol{\Lambda\xi_i} + \boldsymbol{\varepsilon_i}$ with mutually uncorrelated multivariate normally distributed $\boldsymbol{\xi_i}$ and $\boldsymbol{\varepsilon_i}$. Each latent factor $\boldsymbol{\xi_i}$ has a mean and a variance and may correlate with other latent factors $\boldsymbol{\xi_i}$ so that $\boldsymbol{\mu} = \boldsymbol{\Lambda\mu}_{\xi}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}$. Two sets of simulations were done, one dealing with both means and covariance structures as just noted. The other simulation did covariance structure only simulation, estimation, and testing, in which $\boldsymbol{\mu}_{\xi}$, and hence $\boldsymbol{\mu}$, were set at fixed zero vectors.

In both simulations, p = 15 and m = 3, with each factor having five observed indicators. The population parameters are given by

$$\boldsymbol{\Lambda}' = \begin{bmatrix} 0.7 & 0.7 & 0.75 & 0.8 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.75 & 0.8 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.75 & 0.8 & 0.8 \end{bmatrix}$$

and

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & & \\ 0.3 & 1 & \\ 0.4 & 0.5 & 1 \end{bmatrix},$$

with $\text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}$, and thus the unique variances are $\boldsymbol{\Psi} = \mathbf{I} - \text{diag}\left(\boldsymbol{\Lambda\Phi\Lambda}'\right)$. In the structured means model, the factor means are set as $\boldsymbol{\mu}_{\xi} = (1,\ 2,\ 3)'$.

Data generation was accomplished with "lavaan" package (Rosseel, 2012) in R. The data-generating process consists of two steps. We draw common factors $\boldsymbol{\xi_i}$ from a multivariate normal distribution with mean $\boldsymbol{\mu}_{\xi}$ and covariance matrix $\boldsymbol{\Phi}$. Unique factors $\boldsymbol{\varepsilon_i}$ are drawn from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$. These are combined to give observed variables $\boldsymbol{x_i} = \boldsymbol{\Lambda\xi_i} + \boldsymbol{\varepsilon_i}$. This process is repeated N times to obtain one random sample. The simulation studies include sample sizes ranging from 50 to 10,000, which should be enough to examine the performance of the different estimators. The parameter estimates and test statistics from R programming were verified with the EQS software version 6.4 (Bentler, 2006).

In the covariance structure simulation, there are 15 observed variables (p = 15) and 3 latent factors. Thus, $p^* = \frac{15(15+1)}{2} = 120$, with q = 33 free parameters to estimate. The models thus have 87 *df*. For the structured means, there are 15 sample means derived from 15 observed variables and 3 factor means. Altogether, the mean and covariance structure thus has 135 data points and 36 free parameters, yielding 99 *df*. Under the asymptotic properties and multivariate normality, the expected value of the CFA test statistic should be about 87, while its expected value in the mean and covariance structure model should be around 99.

Evaluation criteria

We will focus on two key simulation summaries at different sample sizes: The GLS, ML, and RLS test statistics and their empirical rejection frequencies. Although "lavaan" and EQS can estimate GLS and ML mean and covariance structures, existing R packages cannot compute RLS statistics. Therefore, specialized R code for computing RLS test statistics was written for this study.

The covariance structure tests of interest are T_{ML}, T_{GLS}, and T_{RLS}. Computed in each of 1,000 replications, the empirical means of each of these statistics should be about 87, with an expected standard deviation of $\sqrt{2df} \approx 13.19$. The mean/covariance structure tests are T_{ML_MS}, T_{GLS_MS}, and T_{RLS_MS}, whose empirical mean should be about 99 with an expected standard deviation $\sqrt{2df} \approx 14.07$.

Moreover, p values are the criteria by which the null hypothesis is evaluated with α = 0.05. Each replication will generate a corresponding p value for the fitted model. The average p values of all tests will be calculated. We also use empirical rejection frequencies as one of the benchmarks for evaluating the performance of the test statistics, i.e., the ratio of the number of p values less than 0.05 to the total number of replication (1,000). If the models perform correctly and follow asymptotic properties, their mean rejection rates should be around 0.05 when the sample sizes are sufficiently large. Any deviation far from the α level of 0.05 indicates that the chi-square distribution is not an adequate reference distribution for evaluating model fit.

Results

Test statistics

The left part of Table 1 shows that T_{ML} tends to follow two asymptotic properties when the samples are greater than about 400. First, the test statistics converge to the expected value of 87 as the sample size becomes large, but when $N < 400$, they deviate from the expected value. With $N < 100$, the means are substantially above their expected value. T_{GLS} also follows asymptotic behavior when sample sizes are large; however, when sample sizes are smaller than 500, they are increasingly negatively biased, i.e., behavior opposite to that of T_{ML}. In sharp contrast, the means of T_{RLS} test statistics are highly consistent across all sample sizes – as the sample size varies from 50 to 10,000, the corresponding mean test statistics remain very close to the expected value of 87. These findings are consistent with those of Hayakawa (2019).

The mean and covariance structure test statistics follow similar patterns. The mean of the test statistics T_{ML_MS} and T_{ML_GLS} are near their expected value of 99 with large sample sizes. When sample sizes are smaller, the mean estimates of both T_{ML_MS} and T_{ML_GLS} become increasingly inaccurate, with T_{ML_MS} being excessively large and T_{ML_GLS} excessively small. In contrast, the mean test statistics of T_{RLS_MS} are highly stable across all sample sizes and very close to the expected value of 99.

Figure 1 contrast mean the test statistics of ML and RLS across sample sizes for covariance structures (bottom two lines) and mean/covariance structures (top two lines).

At a glance, the means of T_{ML} and T_{ML_MS} are highly parallel to each other across all samples, as are T_{RLS} and T_{RLS_MS}, although T_{RLS_MS} varies a bit more around its expected value than does T_{RLS}.

The mean standard deviations derived from the 1,000 replications are given in the right part of Table 1. We expect these to be about 13.19 and 14.07 for covariance and mean/covariance structures, respectively, and this is generally found when $N > 400$. However, when sample sizes are smaller, the mean standard deviations of T_{ML} and T_{ML_MS} tend to be larger than those of T_{GLS}, T_{RLS}, and T_{RLS_MS}. The mean standard deviations of T_{RLS} and T_{RLS_MS} tend to be relatively consistent across all sample sizes, that is, these test statistics produce quite stable estimates.

Average p values and empirical rejection rates

The distribution of p values should be uniform under the null hypothesis; hence, the mean p values should be near 0.5. As seen in the left part of Table 2, all methods approximate this at $N = 10{,}000$. T_{ML} tends to have large variation in average p values across different sample sizes, with smaller p values at the smaller sample sizes; the same pattern occurs for T_{ML_MS}. T_{GLS} and T_{GLS_MS} show the opposite pattern, exhibiting average p values that are too high at the smaller sample sizes. In contrast, T_{RLS} and T_{RLS_MS} have mean p values remarkably near 0.5 across all sample sizes.

A more important perspective on the performance of test statistics is given by the p values near the tail of the distribution where accept/reject decisions about models are often made. As the right part of Table 2 shows, when $N = 10{,}000$ the mean empirical rejection rates of all models are near 0.05, so all methods perform well asymptotically. In terms of mean empirical rejection frequency, T_{ML} and T_{ML_MS} share identical patterns. When N is large, they both have about 5% mean rejection rates, but with $N < 400$, the true model is rejected far too frequently (e.g., at $N = 50$, the mean rejection rate is 0.31). T_{GLS}, and T_{GLS_MS} to a lesser extent, has the opposite problem: rejecting the true model too infrequently. In contrast, both T_{RLS} and T_{RLS_MS} have very consistent rejection rates almost across all sample sizes, close to the desired 0.05 level. However, when $N < 200$ T_{RLS_MS} tends to slightly under-reject the true model.

Figure 2 visualizes the rejection rates of ML and RLS statistics across various sample sizes. When $N > 400$ or so, these methods perform similarly, while at smaller Ns, RLS and RLS_MS clearly outperform ML and ML_MS.

Power analysis

In this section, we describe the ability of T_{ML_MS}, T_{RLS_MS}, and T_{GLS_MS} to reject false models, i.e., those that do not correspond to the population that generated the data. If a test statistic requires smaller sample size to reject models with misspecification, then the power of that test is higher. Power analysis is done using three conditions of misspecification.

In condition 1, two extra factor loading parameters are added to the data-generating population model. We connect the second factor with the first manifest variable and third

Table 1. Mean test statistics and standard deviations by sample size.

	Test Statistics						Standard Deviations					
N	ML	GLS	RLS	ML MS	GLS MS	RLS MS	ML	GLS	RLS	ML MS	GLS MS	RLS MS
50	102.32	76.71	87.65	113.80	90.97	98.78	15.53	10.51	12.04	15.68	12.35	12.95
80	96.39	80.70	87.22	106.46	93.62	98.61	14.71	11.78	13.36	14.57	13.45	13.35
100	93.66	82.21	87.10	105.03	94.69	98.33	14.77	12.32	12.55	14.61	13.56	13.74
200	89.86	83.86	87.42	101.88	97.53	98.43	13.99	12.97	13.71	14.00	14.15	13.82
300	89.55	85.50	87.40	100.49	97.74	99.43	14.06	12.65	12.81	14.19	14.33	13.96
400	88.59	85.89	87.31	100.71	98.36	98.23	13.10	12.62	13.16	14.61	14.11	14.13
500	88.54	86.27	87.23	100.67	97.93	98.97	13.76	13.04	12.79	14.41	14.26	13.99
800	87.45	86.20	87.14	99.26	98.02	98.11	13.57	12.68	13.52	13.85	14.00	13.64
1,000	88.06	86.21	86.99	98.90	98.52	98.99	13.51	12.87	13.06	14.45	13.75	13.98
2,000	87.40	87.07	87.13	99.75	99.20	99.12	12.86	12.96	12.87	13.95	13.97	13.66
5,000	87.38	87.12	87.07	99.06	98.79	99.37	13.44	12.94	12.96	13.32	13.50	14.28
10,000	86.58	87.01	86.54	99.43	99.21	99.14	12.95	12.71	13.00	14.31	14.04	14.30

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

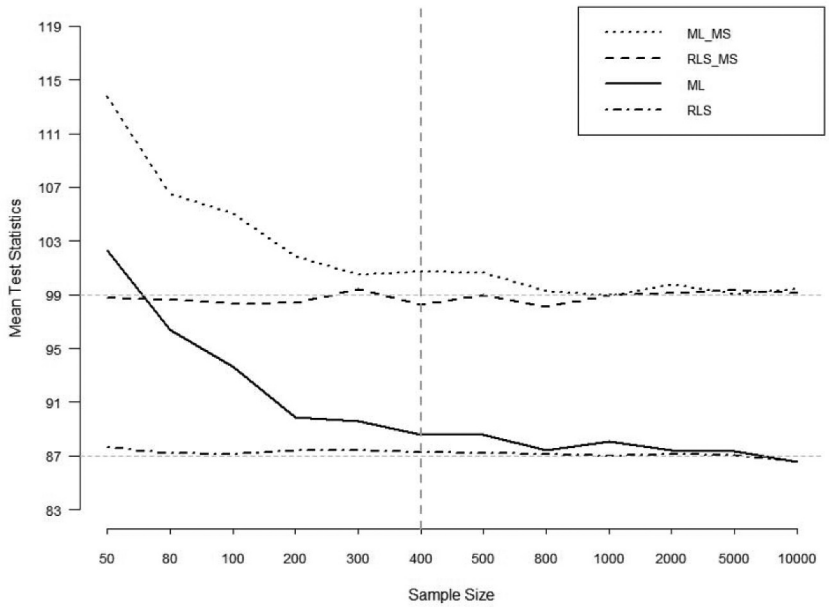

Figure 1. The effect of sample size on mean test statistics.

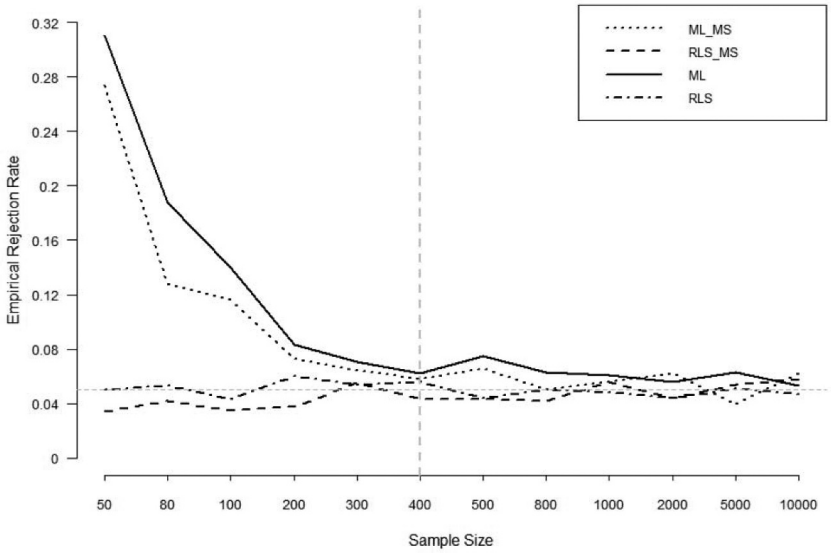

Figure 2. The effect of sample size on empirical rejection frequency.

factor with the sixth manifest variable and set the factor loadings at the values of 0.2 and 0.3, respectively. Thus, the new factor loading matrix is defined as:

.

$$\boldsymbol{\Lambda}' = \begin{bmatrix} 0.7 & 0.7 & 0.75 & 0.8 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.75 & 0.8 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.75 & 0.8 & 0.8 \end{bmatrix}$$

The misspecified model is the one that was defined earlier.

In condition 2, we use original population that omits the two extra parameters, in which the factor means of the population model are 1, 2, and 3. Nonetheless, in the analysis, we fix the factor means at the values of 1, 2, and 2, while holding the intercepts of all other manifest variables at the values of 0. Therefore, there is a misspecification in only one parameter, a latent mean parameter.

Condition 3 is simply a combination of conditions 1 and 2. That is, we analyze the data generated based on the population model specified in condition 1. At the same time, we fix the factor means at the values of 1, 2, and 2. Hence, we expect condition 3 to have a larger misspecification.

Table 2. Simulation results on model *p* values and rejection rates.

	Average P-values						Empirical Rejection Frequencies					
N	ML	GLS	RLS	ML MS	GLS MS	RLS MS	ML	GLS	RLS	ML MS	GLS MS	RLS MS
50	0.226	0.726	0.485	0.240	0.667	0.502	0.310	0.000	0.050	0.274	0.008	0.034
80	0.316	0.638	0.499	0.356	0.610	0.508	0.188	0.009	0.053	0.128	0.020	0.042
100	0.370	0.607	0.497	0.382	0.590	0.510	0.140	0.022	0.043	0.116	0.027	0.035
200	0.441	0.570	0.495	0.441	0.535	0.509	0.083	0.034	0.060	0.073	0.043	0.038
300	0.448	0.531	0.493	0.469	0.528	0.493	0.071	0.036	0.054	0.064	0.049	0.055
400	0.465	0.526	0.495	0.465	0.514	0.515	0.062	0.040	0.056	0.058	0.053	0.043
500	0.470	0.515	0.492	0.465	0.523	0.499	0.075	0.044	0.044	0.066	0.048	0.043
800	0.494	0.515	0.496	0.494	0.518	0.517	0.063	0.039	0.050	0.05	0.042	0.042
1,000	0.479	0.517	0.499	0.503	0.509	0.504	0.061	0.042	0.048	0.056	0.035	0.056
2,000	0.494	0.495	0.494	0.487	0.496	0.496	0.056	0.046	0.044	0.062	0.052	0.044
5,000	0.495	0.497	0.499	0.498	0.502	0.493	0.063	0.049	0.051	0.04	0.045	0.054
10,000	0.512	0.499	0.507	0.493	0.496	0.498	0.053	0.044	0.047	0.062	0.054	0.058

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

As before, for each condition, 1,000 replicated samples were drawn from a population with mean and covariance structure at the previously specified sample sizes. Because the hypothesized models are incorrect, we expect to reject them, and the rejection percentage is an indicator of the power of the test.

The *p* values for each replicated sample were computed, and the mean *p* values and percent of *p* values that are less than $\alpha = 0.05$ are reported in Table 3. This table shows results only for sample sizes between 50 and 1,000 because with $N > 1{,}000$, all rejection percentages were 100.

In all 3 conditions, at all sample sizes, the *p* values for T_{ML_MS} are consistently smaller than those of T_{RLS_MS} and T_{GLS_MS}. With regard to model rejections, in all conditions and at all sample sizes, T_{ML_MS} has the greatest percent rejection as compared to T_{RLS_MS} and T_{GLS_MS}. ML has the most power, showing most clearly that these models are incorrect.

As expected, in all conditions, all tests show increased power to reject the null hypothesis as *N* increases. Condition 1 provides the most challenge to reject the incorrect model, with *N* of about 800 needed for T_{ML_MS} and T_{RLS_MS} to reject the false model 95% or more of the time (T_{GLS_MS} needs a slightly larger *N*). This rejection percent already is achieved at $N = 200$ in conditions 2 and 3 with all methods.

Discussion and conclusion

Scholars in the field of SEM have documented that sample covariance matrix $\boldsymbol{S}$ can be ill-conditioned when sample sizes are small. This has an effect on T_{ML}, specifically, its behavior with the true model is not χ^2 when the sample size is small. Two main solutions have been proposed to remedy this problem: Regularized GLS (RGLS, Arruda & Bentler, 2017) and RLS (Hayakawa, 2019). RGLS is based on Chi and Lange's (2014) MAP covariance matrix estimator, whose basic idea is to replace eigenvalues from a poorly conditioned covariance matrix with shrunken eigenvalues. Arruda and Bentler (2017) have shown that RGLS can produce well-performing test statistics in small samples. This method can be easily extended to estimate mean and covariance structure models. However, the methodology and programming of RGLS is relatively complicated. In contrast, T_{RLS} is much easier to implement and requires less computational power.

Years ago, Harlow (1985) had found that the covariance structure T_{RLS} and T_{ML} perform similarly well when sample sizes are large, but only recently Hayakawa (2019) found that when sample sizes are small, T_{RLS} substantially outperforms T_{ML} in a confirmatory factor model, a panel autoregressive model, and a cross-lagged panel model. The current study affirms this finding. It also shows that, in contrast, the statistical power of T_{RLS} is not as high as that of T_{ML}. We also find that similar patterns hold with mean and covariance structure models. That is, T_{RLS_MS} and T_{ML_MS} perform equally well when the samples are large enough, i.e., both of these methods follow expected asymptotic properties, whereas in the context of small samples, under the true model, T_{RLS_MS} performs better than T_{ML_MS} in terms of chi-square test statistics as shown by empirical rejection frequencies. However, the near-ideal performance of RLS in covariance structures is not fully maintained in mean and covariance structures. That is, with $N < 200$, we found a slight under-rejection in T_{RLS_MS}, i.e., some over-acceptance of the mean/covariance structure. At this time, we do not have a proposal on how to avoid this problem.

With regard to power to reject false models, we found that ML consistently outperforms RLS in both covariance structures and those with structured means. Our conjecture is that this has to do with Equation (8), where greater misspecification would have the effect that the product of $\boldsymbol{S}$ and $\widehat{\boldsymbol{\Sigma}}^{-1}$ does not produce an identity matrix. Hence, the *B* term and hence T_{ML} are larger, increasing power to reject the incorrect model. Of course, with large *N* and/or large misspecification, T_{ML_MS} and T_{RLS_MS} will deliver similar power to reject the incorrect models.

Table 3. Power analysis of T_{ML_MS}, T_{RLS_MS}, and T_{GLS_MS}.

Condition 1						
	ML MS		RLS MS		GLS MS	
N	P-values	Rejection %	P-values	Rejection %	P-values	Rejection %
50	0.20	36.6	0.46	5.3	0.66	1.2
60	0.23	27.9	0.43	7.3	0.61	2.5
70	0.24	27.1	0.42	8.7	0.55	3.1
80	0.25	25.2	0.40	8.2	0.54	4.6
90	0.26	25.3	0.40	10.7	0.52	4.3
100	0.27	23.9	0.37	12.2	0.48	5.1
200	0.21	34.1	0.25	29.4	0.32	15.1
300	0.13	48.1	0.17	37.9	0.21	29.3
400	0.09	59.7	0.10	57.7	0.13	47
500	0.05	76.1	0.06	73.6	0.08	63
800	0.01	96.7	0.01	95.7	0.01	92
1,000	0.00	99.2	0.00	99.4	0.00	99
Condition 2						
	ML MS		RLS MS		GLS MS	
N	P-values	Rejection %	P-values	Rejection %	P-values	Rejection %
50	0.11	54.1	0.32	13.0	0.34	15.3
60	0.11	56.1	0.27	18.5	0.29	20.9
70	0.09	59.3	0.21	26.6	0.23	32.4
80	0.08	62.5	0.18	32.4	0.17	42.6
90	0.07	68.5	0.15	39.9	0.14	49.9
100	0.05	73.7	0.13	46.8	0.11	59.3
200	0.01	97.2	0.01	95.1	0.01	96.8
300	0.00	99.9	0.00	99.9	0.00	99.9
400	0.00	100	0.00	100	0.00	100
500	0.00	100	0.00	100	0.00	100
800	0.00	100	0.00	100	0.00	100
1,000	0.00	100	0.00	100	0.00	100
Condition 3						
	ML MS		RLS MS		GLS MS	
N	P-values	Rejection %	P-values	Rejection %	P-values	Rejection %
50	0.10	58.6	0.29	16.4	0.38	10.1
60	0.09	59.1	0.26	22.5	0.30	21.5
70	0.10	61.0	0.21	29.1	0.23	29.6
80	0.08	64.7	0.17	38.2	0.20	34.0
90	0.07	69.6	0.15	43.6	0.16	42.4
100	0.05	75.2	0.11	50.3	0.13	49.8
200	0.00	97.7	0.01	95.6	0.01	96.5
300	0.00	100	0.00	100	0.00	99.9
400	0.00	100	0.00	100	0.00	100
500	0.00	100	0.00	100	0.00	100
800	0.00	100	0.00	100	0.00	100
1,000	0.00	100	0.00	100	0.00	100

To keep the scope of this paper manageable, its simulations focused on multivariate normal data only. Further research is certainly needed to evaluate the additional complication of test statistics from non-normal distributions using mean and covariance structure models under null and misspecification conditions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Bang Quan Zheng http://orcid.org/0000-0003-2614-2501
Peter M. Bentler http://orcid.org/0000-0002-9440-721X

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). John Wiley & Sons, Inc.

Arruda, E. H., & Bentler, P. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling*, *24*, 657–665. https://doi.org/10.1080/10705511.2017.1318392

Bentler, P. (2006). *EQS 6 structural equations program (Version 6.4)*. Multivariate Software, Inc.

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective*. Wiley-Interscience.

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

Chi, E. C., & Lange, K. (2014). Stable estimation of a covariance matrix guided by nuclear norm penalties. *Computational Statistics and Data Analysis*, *80*, 117–128. https://doi.org/10.1016/j.csda.2014.06.018

Harlow, L. L. (1985). *Behavior of some elliptical theory estimators with nonnormal data in a covariance structure framework: A Monte Carlo study* [PhD Dissertation, UCLA].

Hayakawa, K. (2019). Corrected goodness-of-fit test in covariance structure analysis. *Psychological Methods*, *24*(3), 371–389. https://doi.org/10.1037/met0000180

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

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research*, *26*, 329–367. https://doi.org/10.1177/0049124198026003003

Hu, L.-T., Bentler, P., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362. https://doi.org/10.1037/0033-2909.112.2.351

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, *10*, 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jalal, S., & Bentler, P. (2018). Using Monte Carlo normal distribution to evaluate structural models with nonnormal data. *Structural Equation Modeling*, *25*, 541–557. https://doi.org/10.1080/10705511.2017.1390753

Jöreskog, K. G. (1969). Some contribution to maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. https://doi.org/10.1007/BF02289343

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons Ltd.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, *19*, 86–98. https://doi.org/10.1080/10705511.2012.634724

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

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

Yuan, K. H., & Bentler, P. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774. https://doi.org/10.1080/01621459.1997.10474029

Yuan, K. H., Zhang, Z., & Deng, L. (2019). Fit indices for mean structures with growth curve models. *Psychological Methods*, *24*, 36–53. https://doi.org/10.1037/met0000186