Review Article

Structural Equation Modeling as a Statistical Method: An Overview

Ke-Hai Yuan1* and Yubin Tian2
1Department of Psychology, University of Notre Dame, USA
2Department of Applied Mathematics, Beijing Institute of Technology, China

Abstract

Important attributes in many disciplines cannot be observed directly, and have to be measured by multiple indicators that are subject to errors. Structural equation modeling (SEM) has been a major tool for understanding the relationships between latent attributes and their observed indicators as well as those among latent attributes. This article provides a critical review of the statistical development of SEM methodology. Pros and cons of each method are noted, and the need for further development is pointed out. In particular, there is no effective procedure for SEM with small N and/or large p, together with incomplete data from an unknown population distribution.

ABBREVIATION

AOSEM: An Overview of Structural Equation Modeling.

INTRODUCTION

In social, behavioral, education and health sciences, important attributes are often latent variables that cannot be observed directly, and have to be inferred from fallible measures. Their analyses are most effectively done by structural equation modeling (SEM). In contrast to statistical methods in traditional multivariate analysis, SEM has the mechanism of modeling manifest variables, latent variables, as well as measurement errors simultaneously. Although not familiar to many statisticians, SEM has become one of the most important methodology in many disciplines when analyzing survey or non- and quasi-experimental data. The journal, Structural Equation Modeling, launched in 1994, is consistently ranked as having the highest impact among all statistical and mathematical-method journals (http://en.wikipedia.org/wiki/Comparison_of_statistics_journals).

A SEM model can be equivalently represented as a mean and covariance structure model, and many commonly used statistical models are its special cases. The confirmatory factor model is among most widely used SEM models [1-2]. Other popular SEM models include the latent growth curve model [3-4], multitrait-multimethod model [2], and multiple-indicator and multiple-cause model [2,5]. In particular, regression, MANOVA, and path analysis are also special cases of SEM models [6-7]. Commercial software for SEM includes AMOS, EQS, LISREL, MPLUS, and SAS CALIS; and free software includes Mx, sem and lavaan in R, and WebSEM.

The purpose of this article is to provide an overview of SEM methodology, include estimation methods, test statistics, and different formulations of standard errors of parameter estimates. Pros and cons of each method are noted, and the need for further research is pointed out.

ESTIMATION METHODS

Like in any other disciplines of statistics, estimation methods in SEM were mostly developed for obtaining more efficient parameter estimates, catering for different types of data and/or distributions in practice. Among all the methods of SEM, the normal-distribution-based maximum likelihood (NML) [8] is most widely used. Let \( \mathbf{T} \) and \( \mathbf{S} \) be the sample means and sample covariance matrix, and the corresponding mean and covariance structure model be \( \mu(\theta) \) and \( \Sigma(\theta) \). The NML method obtains parameter estimate \( \hat{\theta} \) by minimizing the discrepancy function

\[
F_{\text{NML}}(\theta) = [\mathbf{T} - \mu(\theta)]\Sigma^{-1}(\theta)[\mathbf{T} - \mu(\theta)] + \text{tr}[\Sigma^{-1}(\theta)] - \log|\Sigma^{-1}(\theta)| - p, (1)
\]

Where \( p \) is the number of observed/manifest variables. Under the assumption of normally distributed data, minimizing the \( F_{\text{NML}}(\theta) \) in (1) yields asymptotically most efficient parameter estimates. However, normally distributed data are as rare as unicorn in practice [9]. In particular, NML estimates are inefficient or biased when samples are from heavy-tailed population distributions or contain outliers.

Various alternative methods have been developed to account for nonnormality with real data. The most well-known development is the Generalized Least Squares (GLS) procedure [10,11], which is commonly referred to as the asymptotically distribution free (ADF) method. Let \( s = \text{vech}(S) \) be the vector...
containing the non-duplicated elements of \( S \), \( \sigma(\theta) \) be the model counterpart of \( s \), \( v(\theta) \) be \( \mu(\theta) \), \( \sigma(\theta) \). Then, \( \hat{\Gamma} = \Sigma_{\mu\Sigma} / N = (\bar{y}, y' \Sigma y) / (N, y' y) \). Let \( \Gamma \) be the asymptotic covariance matrix of \( \sqrt{N} \). Parameter estimate \( \hat{\theta} \) in the GLS/ADF procedure is obtained by minimizing the discrepancy function

\[
F_{GLS}(\theta) = \sqrt{\hat{\Gamma} - [\hat{y}' - v(\theta)]} \hat{\Gamma}^{-1} \sqrt{\hat{y}' - v(\theta)},
\]

where \( \hat{\Gamma} = \Sigma_{\hat{y}} \) is the sample covariance matrix of \( y_i \), \( i = 1, 2, \ldots, N \). Under the condition that the population 4th-order moments exist, the GLS estimate \( \hat{\theta} \) is asymptotically most efficient among all methods of modeling \( \mu \) and \( \Sigma \) via \( \hat{\bar{y}} \) and \( \hat{\Sigma} \). However, Monte Carlo results indicate that, unless the sample size \( N \) is rather large, GLS estimates are not as efficient as NML estimates even when the population distribution substantially deviates from \( \mathcal{M}(\mu, \Sigma) \) [12]. Special cases of GLS are the least squares (LS) method when \( \hat{\Gamma} = I \) (the identity matrix) and the normal-distribution-based GLS for which the \( \hat{\Gamma} \) in (2) is replaced by the normal-distribution-based 4th-order moment matrix.

Notice that the GLS function in (2) does not use any particular distribution shape of the underlying population. Methods accounting for the distribution shape of the underlying population include the ML method based on an elliptical distribution [13-15] and one that allows each marginal distribution to have its own form of 4th-order moment [16]. These methods would perform ideally when the underlying population distribution is as assumed. However, it is not clear how they perform when the assumption is violated.

Another class of SEM methods accounting for heavy-tailed distributions in general is via robust estimates of \( \mu \) and \( \Sigma \). Robust \( \hat{\mu} \) and \( \hat{\Sigma} \) can be obtained by M-estimates using ML based on a multivariate \( t \)-distribution or estimating equations based on Huber-type weights [17]. With a corresponding consistent \( \hat{\Gamma} \), robust SEM can be done by replacing the \( \hat{\bar{y}} \) and \( \hat{\Sigma} \) in (1) or (2) together by \( \hat{\bar{y}} \) by their robust counterparts [18,19]. With real data, the robust methods have been shown to yield much more reliable estimates than NML and/or GLS [20,21]. However, these methods have not been widely used because applied researchers are not aware of the merit of robust methods or do not have access to software [6].

Other methods of SEM have been developed for special data structures, including the GLS or ML method for ordinal data by modeling the polychoric correlation matrix [22,23]; GLS and ML methods for samples with missing data [24], and robust methods for missing data with auxiliary variables [25].

Notice that the merit of ML, GLS, or any of the other methods reviewed so far is justified by asymptotics. In practice, obtaining a sample with a sufficiently large \( N \) may be unrealistic, because most researchers only have limited resources in data collection. In particular, when the number of items \( p \) is large, we may have difficulty in getting a positive definite matrix \( S \) or a positive definite \( \hat{\Gamma} \). Then the function \( F_{ML}(\theta) \) in (1) or \( F_{GLS}(\theta) \) in (2) is not defined. When \( S \) is close to being singular, the program LISREL [26] provides an option of ridge SEM by replacing the \( S \) in equation (1) with \( S + k \mathbf{diag}(s_{11}, \ldots, s_{pp}) \), where \( k > 0 \) and \( s_{ij} \) is the sample variance of the \( j \)th variable. However, statistical properties of the resulting parameter estimates and related test statistics are not as desired [27].

A different ridge procedure was developed by Yuan and Chan [28], who proposed to replace the \( S \) in equation (1) by \( S + aI \) and recommended choosing \( a \approx p / N \), where \( I \) is the identity matrix. Let the resulting discrepancy function be denoted by \( F_{ML}(\theta) \) and call the procedure of minimizing \( F_{ML}(\theta) \) for \( \hat{\theta} \) ridge ML. Yuan and Chan [28] showed that ridge ML yields consistent and more accurate parameter estimates than NML even when data are normally distributed. Yuan et al. [29] further proposed of modeling the polychoric/polyserial correlation matrix with ordinal data by ridge ML, and showed that the resulting parameter estimates are more efficient than other existing methods.

Recently, Yuan and Chan [30] proposed to replace the \( \hat{\Gamma} \) in (2) by \( (1-a)S + aI \) and called the procedure ridge GLS, where \( a \in [0,1] \) is a tuning parameter. Simulation results indicate that, for nonnormally distributed data, ridge GLS can yield much more efficient parameter estimates than NML, ridge ML, and GLS.

In summary, there are a variety of methods for estimating model parameters in SEM. Results in real data analyses as well as in Monte Carlo studies indicate that robust methods or ridge GLS methods tend to yield most efficient or accurate parameter estimates [30,31]. In particular, robust methods with missing data and auxiliary variables have been implemented in a window-based program WebSEM that is freely accessible through a web browser [32].

**TEST STATISTICS**

Most SEM models with latent variables are rather hypothetical even if they are supported by substantive theory. Thus, model evaluation/testing is a key component of SEM methodology. Different model test statistics have been developed under different estimation methods, more than one statistic might also be constructed for a given set of parameter estimates. This section describes existing developments of test statistics in SEM. Most of the reviewed statistics are originally developed for only covariance structure model \( \Sigma(\theta) \), which can be regarded as a special case of mean and covariance structure model \( \mu(\theta), \Sigma(\theta) \) with a saturated mean structure. We will not separately discuss these two kinds of models to save space.

The classical test statistic for overall model evaluation, corresponding to the NML parameter estimator, is \( T_{ML} = (N-1)F_{ML}(\hat{\theta}) \), which is essentially the likelihood ratio statistic under the normality assumption. \( T_{ML} \) is the default statistic in most SEM software and also most widely used in practice. With normally distributed data and a sufficiently large \( N \), \( T_{ML} \) approximates follows the nominal chi-square distribution \( \chi^2_q \), where \( df = p(p+3)/2 - q \) with \( q \) being the number of free parameters in \( \mu(\theta) \) and \( \Sigma(\theta) \). However, \( T_{ML} \) is strongly affected by violation of normality and data contamination. Also, with normally distributed data, the requirement on \( N \) for \( T_{ML} \) to approximately follow \( \chi^2_q \) is increasingly demanding as 1

---

1 There exist multiple studies for \( \chi^2_{ML} \) to asymptotically follow a chi-square distribution for nonnormally distributed data, called asymptotic robustness. Because the conditions behind asymptotic robustness cannot be effectively verified, we will not discuss such literature in this review.
$p$ increases. Results in Hu, Bentler and Kano [33] indicate that, for data from heavy-tailed distributions, $T_{uc}$ can reject correct model 100% when the nominal level is 5%. Similarly, results in Moshagen [34] with normally distributed data indicate that, with $N = 200$ and nominal type I error rate being set at 5%, $T_{uc}$ rejects the correct models 98% at $p = 60$ and 100% at $p = 80$.

There are multiple statistics that account for nonnormality of the underlying population distribution. The most widely used one is the rescaled statistic $T_{RMLT} = (df / \hat{c}_T)T_{ML}$, proposed by Satorra and Bentler [35], where $\hat{c}_T$ is a consistent estimate of $E(T_{ML})$ by combining model structure $\mu(\theta)$ and $\Sigma(\theta)$ as well as the 3rd- and 4th-order moments of the sample. In particular, as $N \longrightarrow \infty$, $T_{RMLT}$ approaches a distribution whose mean equals $df$. Another is an adjusted statistic $T_{AMLT} = \hat{c}_T^2T_{ML}$, where $\hat{c}_T$ also accounts for model structure as well as sample 3rd- and 4th-order moments so that both the mean and variance of $T_{AMLT}$ approach those of the corresponding reference chi-square distribution. Among the two, $T_{RMLT}$ is more well-known and is called a robust chi-square statistic by some authors. Monte Carlo results indicate that $T_{RMLT}$ and $T_{AMLT}$ perform reasonably well for medium to large $N$ [37,38]. However, there exists evidence that they do not work well with small $N$ and/or large $p$ [37,38]. In fact, $T_{RMLT}$ or $T_{AMLT}$ may not be properly defined when $N \leq p = p(p+3)/2$ [37], and $N$ needs to be substantially greater than $p^3$ for them to work reasonably well. Also, since both $T_{RMLT}$ and $T_{AMLT}$ use the NML estimate, they are not robust to data contamination or outliers [18].

Another well-known statistic accounting for nonnormality is the GLS estimator $\hat{\theta}$. As long as the underlying population distribution has finite 4th-order moments, the statistic $T_{GLS} = (N-1)F_{GLS}(\hat{\theta})$ asymptotically follows a chi-square distribution. However, empirical results in Hu et al. [33] indicated that, at $p=15$ and $N=150$, $T_{GLS}$ rejects the correct model 100% when the target is 5%.

Following the GLS parameter estimate of minimizing (2), statistics accounting for both nonnormality and small $N$ include a corrected GLS statistic $T_{CGLS} = T_{GLS}/(1 + F_{GLS}(\hat{\theta}))$ [39] and an $F$-statistic $T_F$ [40]. In parallel, corrected GLS and $F$ statistics have also been developed following the MLM estimation procedure through the vector $\tau - \nu(\hat{\theta})$ of residuals, called corrected residual-based GLS/ADF statistic [41], where $\hat{\theta}$ is the NML estimate. Although these statistics perform reasonably well for medium to large sample sizes [36,38], when $N$ is relatively small, $T_F$ tends to over-reject the correct models and $T_{CGLS}$ tends to under-reject the correct models [37]. In particular, like $T_{RMLT}$ and $T_{AMLT}$, $T_F$ and $T_{CGLS}$ or their residual-based counterparts may not be properly defined when $N \leq p^3$.

Corresponding to robust estimation methods with complete or incomplete data, rescaled and adjusted statistics parallel to $T_{RMLT}$ and $T_{AMLT}$ have been developed [18,25]. Monte Carlo results indicate that these statistics perform more reliably than their non-robust counterparts [42].

Similarly, rescaled and adjusted statistics following ridge ML estimation method work substantially better than their NML counterparts [28,29]. Rescaled and adjusted statistics following ridge GLS also perform substantially better than $T_{CGLS}$ [30].

There are also a variety of developments for improving test statistics with small $N$ under the assumption of a normally distributed population, around replacing $(N-1)$ in $T_{ML} = (N-1)F_{ML}(\hat{\theta})$ by a smaller number so that the resulting statistic $T$ approximately satisfies

$$E(T | H_0) = df,$$

and consequently $T$ follows $\chi^2_{df}$ more closely. Pioneer studies in this area include Bartlett [43,44], Box and Lawley [45,46]. The following developments in this direction are called Bartlett correction [47]. However, these corrections are too complicated to apply for general SEM models.

In the context of exploratory factor analysis (EFA), the normal-distribution-based likelihood ratio statistic can also be written as $T_{ML} = (N-1)F_{ML}(\hat{\theta})$, where $F_{ML}(\theta)$ is given by (1). Bartlett [48] proposed a simple correction to replace $N-1$ in $T_{ML}$ in the context of EFA by $N = (N-2p+1)/(6 - 2m/3$ [49], where $m$ is the number of latent factors. Because the resulting statistic, $T_{MLc}$, is easy to implement, Fouladi [36] proposed to apply the simple correction to SEM models with $m$ remaining to be the number of latent factors. However, results in Nevitt and Hancock [38] indicate that type I errors with $T_{MLc} \sim \chi^2_{df}$ in SEM tend to be much lower than the nominal level. Considering that the number of parameters in SEM increases at a much lower rate than that in EFA when $m$ increases, Yuan (2005) [50] proposed to replace $(N-1)$ in the definition of $T_{ML}$ by $N = (N-2p+13)/6 - m/3$. A more complicated correction was offered by Swain (1975) [51], who proposed to replace $(N-1)$ in $T_{ML}$ by

$$N = N - 1 - \{p(2p^2 + 3p - 1) - h(v_{20} + 3h_{q-1})/12(df),$$

where $h_i = (1+8q)^{i/2}/2$ and $q$ is the number of free parameters in $\theta$. Studies by Fouladi [36] and Herzog and Boomsma [52] indicate that the statistics performing from best to worst are $T_{MLb} = N, F_{ML}(\theta), T_{MLy} = N, F_{ML}(\hat{\theta}), T_{MLb,},$ and $T_{ML}$. Although the performance of $T_{MLc}$ is potentially promising, the correction factor $N$ is not statistically justified.

Yuan, Tian and Yanagihara [53] proposed an empirical modeling approach to correct $T_{MLc}$ so that the resulting statistic approximately satisfies (3). Let

$$\Delta = 1 - (p^2 + p\hat{\beta} + m\hat{\beta} + q\hat{\beta}_y)/(N-1),$$

where the $\hat{\beta}$s are parameters to be estimated by fitting (3) with $T = \Delta T_{MLc}$ and simulated values of $T_{ML}$. With many replications, $\Delta = \Delta T_{MLc}$ can be regarded as approximately following a normal distribution, and $\beta$ can thus be estimated by normal-distribution-based ML. With 342 different model and data conditions and 200 or 500 replications for each condition, one of the empirically corrected statistics in Yuan, Tian and Yanagihara [53] is

$$T_{MLb} = N - 1 - \{p(2.381 + 3.677p + 0.003q_2)\}F_{ML}(\hat{\theta}),$$

where $q_2 = q - 2p$ is the number of parameters beyond that of a one-factor model. According to the size of standardized residuals and the Bayesian information criterion (BIC), $T_{MLb}$ performs much better than the analytically corrected ones in fitting (3), including $T_{MLc}$. Yuan, Tian and Yanagihara [53] also conducted an independent Monte Carlo study examining type I errors of these statistics with models distinct from any of the 342 conditions used in calibrating $T_{MLc}$, and found that $T_{MLb}$ performs...
significantly better than $T_{MLeT}$ and other analytically corrected ones. In particular, $T_{MLeT}$ controls type I errors reasonably well when $N \geq 2p$. However, the applicability of the $T_{MLeT}$ in (4) is only for normally distributed data, and practical data are typically nonnormally distributed.

In summary, a lot of efforts have been made towards obtaining reliable test statistics for overall model evaluation. But none of them controls type I errors desirably when $N$ is small and $p$ is large, especially for samples with incomplete data from unknown population distributions. More developments are needed to yield test statistics that work for both nonnormally distributed data and small $N$ with a large $p$.

### STANDARD ERRORS (SEs)

The classical method for calculating the SEs of NML parameter estimates is by inverting the corresponding information matrix. With real data whose population distributions are typically unknown, such SEs are not consistent in general and thus not valid even asymptotically. Consistent covariance matrices for NML estimates can be obtained through the so-called sandwich-type covariance matrices [10,13,54] and they have been implemented in standard software (e.g., EQS, lavaan in R, Mplus, WebSEM). The corresponding sandwich-type SEs are more reliable than the NML-based SEs when $N$ is relatively large. But sandwich-type SEs do not perform well when $N$ is small to medium [12].

Formulas of sandwich-type covariance matrix have been developed for SEM with samples from heavy-tailed distributions via robust $\hat{\mu}$ and $\hat{\Sigma}$ [18,19], with samples containing missing data [24,25], and for ridge ML with continuous and ordinal data [28,29] as well as ridge GLS method [30]. Again, all these proposals are justified by asymptotics, and fine tuning is needed for the sandwich-type SEs to work well with small $N$.

According to asymptotics, the covariance matrix of the GLS estimator [10] is given by $(\psi^T \psi)^{-1} / (N - 1)$, where $\psi$ is the matrix of partial derivatives of $v(\theta)$ with respect to $\theta$. However, when evaluated at the GLS estimator $\hat{\theta}$ and $\hat{\psi} = \hat{S}$, the corresponding SEs are simply too small when compared to the empirical ones [12]. SEs based on the corrected $(\psi^T \psi)^{-1} / (N - p^2 - 2)$ are still too small when $N$ is not large enough. More work is needed for reliable SEs of the GLS estimator.

### OTHER PROPOSALS

In addition to methods reviewed above, other methods have been proposed to address various issues in SEM, including bootstrap [55,56], Bayesian method [57], and two-stage least squares (2SLS) with instrumental variables [58]. We will briefly discuss the pros and cons of these methods below, and leave the details to references where these methods were studied.

For any of the reviewed estimation methods, the bootstrap method yields consistent standard errors of the corresponding parameter estimates [59] and consistent estimate of type I errors for overall model evaluation [60]. When the sample size is not too small, bootstrap also yield reliable inferences. However, the bootstrap method is computationally intensive, and may run into problems when the percentage of non-converged replications is substantial [56,61]. With bootstrap, each replication is a subset of the original sample with duplicated cases, and hence non-convergence can be a serious problem with a small $N$.

The Bayesian method also yields reliable inference when the conditional distribution of the observed data given $\theta$ can be correctly specified. In particular, the method allows researchers to include prior information. However, with typical nonnormal data from unknown population distributions in social and behavioral sciences [9], Bayesian methods may not lead to consistent analysis. Also, with many tuning elements in the process of Markov chain Monte Carlo (e.g., prior distribution, the number of burning iterations, and the way to evaluate convergence of different Markov chains), it is an art to effectively use the Bayesian methodology.

In 2SL approach, identification of instrumental variables is based on the hypothetical model structure. With correct identification of instrumental variables, misspecifications in one part of the model may not affect the estimation of the other part of the model. However, when one is not aware of the misspecification in a theoretical model, the 2SL approach can yield biased estimates of factor loadings and other parameters as well [62].

### DISCUSSION AND CONCLUSION

Structural equation modeling has become one of the most important statistical methods in analyzing survey or non-experimental data. The methodology has found wide applications in social, behavioral, education, management, and health sciences, where errors in measurements cannot be avoided. As reviewed in the previous sections, many methods have been developed for parameter estimation and model inference. However, much more work is needed to advance the methodology of SEM. In particular, with typical data from unknown population distributions, small sample size $N$ and large number of items $p$, together with incomplete data, reliable statistical inference in SEM is rather challenging.

We only reviewed the key statistical components of SEM to keep the article focused. Other important but less statistical components of SEM include model identification (Bollen, 1989, pp. 88–104), computation algorithm [63] and fit index [50,64]. Also, we only reviewed the components of basic SEM models with a homogeneous population. More complicated SEM models have also been developed for samples from multiple groups [65-67], samples with a hierarchical structure [68-70], and samples from heterogeneous populations with group memberships unknown a priori [71,72]. A detailed overview of the technical developments of SEM prior to 2007 is given by Yuan and Bentler [73].

### REFERENCES


