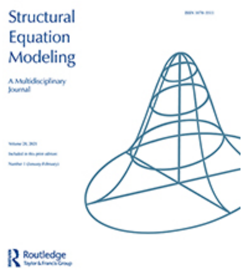

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

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

Which Method is More Reliable in Performing Model Modification: Lasso Regularization or Lagrange Multiplier Test?

Ke-Hai Yuan & Fang Liu

To cite this article: Ke-Hai Yuan & Fang Liu (2021) Which Method is More Reliable in Performing Model Modification: Lasso Regularization or Lagrange Multiplier Test?, Structural Equation Modeling: A Multidisciplinary Journal, 28:1, 69-81, DOI: 10.1080/10705511.2020.1768858

To link to this article: https://doi.org/10.1080/10705511.2020.1768858

Published online: 09 Jul 2020.

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2021, VOL. 28, NO. 1, 69–81
https://doi.org/10.1080/10705511.2020.1768858

Which Method is More Reliable in Performing Model Modification: Lasso Regularization or Lagrange Multiplier Test?

Ke-Hai Yuan [1] and Fang Liu [2]

[1]University of Notre Dame and Nanjing University of Posts and Telecommunications; [2]Renmin University of China

ABSTRACT
Data-driven model modification plays an important role for a statistical methodology to advance the understanding of subjective matters. However, when the sample size is not sufficiently large model modification using the Lagrange multiplier (LM) test has been found not performing well due to capitalization on chance. With the recent development of lasso regression in statistical learning, lasso regularization for structural equation modeling (SEM) may seem to be a method that could avoid capitalizing on chance in finding an adequate model. But there is little evidence validating the goodness of lasso SEM. The purpose of this article is to examine the performance of lasso SEM by comparing it against the LM test, aiming to answer the following five questions: (1) Can we trust the results of lasso SEM for model modification? (2) Does the performance of lasso SEM depend more on the effect size or the absolute value of the parameter? (3) Does lasso SEM perform better than the widely used LM test for model modification? (4) Are lasso SEM and LM test affected by nonnormally distributed data in practice? and (5) Do lasso SEM and LM test perform better with robustly transformed data? By addressing these questions with real data, results indicate that lasso SEM is unable to deliver the expected promises, and it does not perform better than the LM test.

KEYWORDS
Structural equation modeling; real data; robust transformation; power; type I error

Introduction

Due to the capability of modeling measurement errors and latent traits simultaneously, structural equation modeling (SEM) has become one of the most widely used methods for analyzing survey and quasi-experimental data. In applications of SEM, researchers typically begin with an initial model and use a computer program to fit the model to empirical data. If the fit of the initial model is considered inadequate, then it is tempting to use the results of the Lagrange multiplier (LM) test to modify the model to improve its fit to the data (Bentler, 2006, Sörbom, 1989). However, the study of MacCallum et al. (1992) indicates that data-driven model modification is inherently susceptible to capitalization on chance, especially when the sample size is not sufficiently large. With the recent development of regularized regression in statistical learning (Tibshirani, 1996), lasso (least absolute shrinkage and selection operator) SEM may seem to be the method for finding a sound model with an adequate fit (see e.g., Huang et al., 2017; Jacobucci et al., 2016; Pan et al., 2017). However, the literature contains little evidence in validating the goodness of lasso SEM. The purpose of this article is to examine the performance of lasso SEM by comparing it against LM test. In particular, by working with real data, we are going to examine whether lasso SEM performs more reliably than the LM test in locating the omitted parameters for an initial model that does not fit the population adequately.

Data-driven model modification plays an important role for a statistical methodology to advance the understanding of subjective matters. Model diagnostics and techniques of variable selection in regression analysis are post-hoc methodology. To minimize the risk of model modification, criteria such as AIC, BIC, and cross-validation in model selection have been widely used in regression analysis and other fields of statistics. These criteria have been adopted to SEM and covariance structure analysis. Although their routine applications are not risk-free, few better alternatives are available in statistical modeling with empirical data.

In regression analysis, when the number of variables is large and sample size is not sufficiently large, lasso regularization has been shown to be a viable method for variable/model selection. In particular, lasso regression with its adaptive version (i.e., MCP, SCAD) has been shown to possess the so-called oracle property (Fan & Li, 2001; Zou, 2006). That is, the methodology automatically selects the true model if there exists one. Because of the promises of lasso methodology, it has been extended to factor analysis and SEM (Choi et al., 2010; Hirose & Yamamoto, 2014, 2015; Jacobucci et al., 2016; Ning & Georgiou, 2011; Pan et al., 2017). Huang et al. (2017) showed that lasso SEM and its extension possess the oracle property. However, they also pointed out that the oracle property is based on asymptotics that may not be realizable with finite samples. Clearly, like LM test, lasso SEM is also a data-driven methodology. According to MacCallum et al. (1992), "any strategy that is data-driven is inherently susceptible to problems arising from capitalization on chance." Although lasso SEM provides a lot of promises, there is little evidence showing that it is free from capitalization on chance. A study is needed to evaluate the actual performance of the method before it can be routinely used with real data.

CONTACT Ke-Hai Yuan kyuan@nd.edu Department of Psychology, University of Notre Dame, Notre Dame, IN 46656, USA; Fang Liu fang_liu_h@163.com School of Statistics, Renmin University of China, Beijing 100872, China.

While the modification index/LM test has been shown to be affected by idiosyncratic characteristics of the sample, results in MacCallum et al. (1992) and Green and Thompson (2010) also indicate that LM test performs quite well at relatively large sample sizes. The literature on lasso SEM contains results with simulated normally distributed data (e.g., Pan et al., 2017). But there is little direct comparison between lasso SEM and LM test. Also, because the discrepancy function of lasso SEM or Bayesian lasso SEM is derived from normally distributed data, it is not clear whether the findings with simulated normal data are over-optimistic when the methodology is applied to real data. As applications of lasso SEM, examples with real data show that the method performs reasonably well (e.g., Huang et al., 2017; Jacobucci et al., 2016). However, because we do not know the true model with real data, it is hard to tell to what degree the oracle property of lasso SEM is realized with finite samples. Similarly, the method of model modification examined by MacCallum et al. (1992) is based on the LM test derived from a normality assumption but applied to real data. It is also not clear whether the poor performance of the LM test is due to the LM test itself or due to the violation of normality with the real data. In particular, results in Yuan et al. (2002) indicate that the cross-validation index derived from normally distributed data is severely affected by nonnormality, and robust methods can effectively address the problem. In this article, we also aim to clarify these issues in addition to directly compare LM test and lasso SEM.

We will compare the performances of LM test and lasso SEM using real data whose population distribution is unknown. For studying the oracle properties of lasso SEM, we will further transform the data so that a correct covariance structure model exists for the population and is known. A robust-transformation technique is used to deal with the effect of nonnormality, and lasso SEM and LM test are compared both before and after the transformation. Via these manipulations we aim to answer the following five questions: (1) Can we trust the results of lasso SEM for model modification? (2) Does the performance of lasso SEM depend more on the effect size or the absolute value of the parameter? (3) Does lasso SEM perform better than the widely used LM test or model modification index in standard SEM software? (4) Are lasso SEM and LM test affected by nonnormally distributed data in practice? and (5) Do lasso SEM and LM test perform better with robustly transformed data?

We are going to compare the analytical features of lasso SEM and LM test in the next section. A real data set is introduced in a following section together with the technique of robust transformation. Methods for conducting the data analysis are described in a subsequent section. Results and analysis of the comparison are presented in a separate section. Recommendation and discussion will be provided in the concluding section.

Lagrange multiplier test and lasso regularization

In this section, we compare the formulations of LM test and lasso SEM, by discussing their similarity and differences. While there are technical developments behind each method with respect to computation and asymptotic properties, we are not going to review these rather technical details. Interested readers are referred to Huang et al. (2017), Huang (2020), and Jacobucci et al. (2016) for lasso SEM, and to Lee and Bentler (1980), Bentler and Dijkstra (1985) and Sörbom (1989) for LM test in SEM, and to Silvey (1959) and Rao (1948) for LM and score test more broadly.

Both LM test and lasso SEM can be carried out with any distribution-based likelihood or discrepancy function. In the literature of SEM, they are commonly operated via the normal-distribution-based maximum likelihood (NML). As is well known, except for using an unbiased sample covariance matrix, the NML method for covariance structure analysis is equivalent to minimizing the discrepancy function

$$F_{ml}(\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta})] - |\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})| - p,$$

where $\mathbf{S}$ is the sample covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the structural model, and p is the number of observed variables. For a typical specification of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represented by a path diagram, there are many zero values of parameters symbolized by no direct path between variables/factors. In particular, an omitted path can be represented by the constraint $\theta = 0$, where θ is the value of the path coefficient; and two paths with equal coefficients can be represented by the constraint $\theta_1 - \theta_2 = 0$. LM test and lasso SEM have the same purpose, i.e., to identify the omitted paths that are responsible for lack of fit in practice. Instead of directly minimizing $F_{ml}(\boldsymbol{\theta})$, they each work with an augmented function

$$Q(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F_{ml}(\boldsymbol{\theta}) + g(\boldsymbol{\lambda}, \boldsymbol{\theta}), \tag{1}$$

where $g(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is the augmented component according to specific constraints on model parameters, and $\boldsymbol{\lambda}$ is a vector of Lagrangian-multiplier parameters introduced for implementing the constraints (https://en.wikipedia.org/wiki/Lagrange_multiplier).

The difference between lasso SEM and LM test lies in the formulation of $g(\boldsymbol{\lambda}, \boldsymbol{\theta})$. For a set of constraints represented by $\mathbf{c}(\boldsymbol{\theta}) = (c_1(\boldsymbol{\theta}), c_2(\boldsymbol{\theta}), \ldots, c_h(\boldsymbol{\theta}))' = \mathbf{0}$, the formulation of $g(\boldsymbol{\lambda}, \boldsymbol{\theta})$ in LM test is given by

$$g(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_h)'$. The form of $\mathbf{c}(\boldsymbol{\theta})$ for omitted paths is very simple in the context of model modification. Let $\boldsymbol{\gamma}$ be the set of free parameters in the initial model, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_h)'$ be the set of parameters that are fixed at zero in the initial model but freeing them will potentially improve the fit of the model, and $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\nu}')'$. Then, $\mathbf{c}(\boldsymbol{\theta}) = \boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_h)'$ and the augmented component for LM test is

$$g(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{j=1}^{h} \lambda_j \nu_j. \tag{3}$$

We will need to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ with the augmented function $Q(\boldsymbol{\theta}, \boldsymbol{\lambda})$ of the LM test. They can be obtained by solving the equations

$$\partial Q(\boldsymbol{\theta}, \boldsymbol{\lambda})/\partial\boldsymbol{\theta} = \mathbf{0} \text{ and } \partial Q(\boldsymbol{\theta}, \boldsymbol{\lambda})/\partial\boldsymbol{\lambda} = \mathbf{0}. \tag{4}$$

In particular, the principle of constrained maximum likelihood also yields an asymptotic covariance matrix and standard errors for $\hat{\boldsymbol{\lambda}}$ (Bentler & Dijkstra, 1985; Silvey, 1959). These are then used to construct tests of whether an individual or a subset of the $\hat{\lambda}_j$ is statistically significant. If so, then the corresponding constraint $v_j = 0$ needs to be released, and the value $(\hat{\lambda}_j/\widehat{\mathrm{SE}}_j)^2$ is the expected drop in the likelihood ratio statistic $T_{ml} = (N-1)F_{ml}$. Thus, the LM test is a deterministic procedure, and except the significance level (e.g., .05) there is no tuning parameter in its implementation. We will use the LM test as implemented in EQS (Bentler, 2006) to conduct the analysis in the following sections.

For the same set of parameters omitted from the initial model, the augmented component of lasso SEM is

$$g(\lambda, \boldsymbol{\theta}) = \lambda \sum_{j=1}^{h} |v_j|, \tag{5}$$

where the Lagrangian multiplier is a scalar (i.e., $\boldsymbol{\lambda} = \lambda$) corresponding to a composite inequality constraint $\sum_{j=1}^{h} |v_j| \le c$. Unlike the Lagrangian multipliers in the LM test, the value of λ in lasso SEM is selected first and then the function $Q_\lambda(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \lambda)$ is minimized. Clearly, $\hat{\boldsymbol{\theta}}$ will be the NML estimates when $\lambda = 0$, and all the $\hat{v}_j$ need to be zero when λ is sufficiently large. As λ varies, a certain set of the $\hat{v}_j$ will be zero in order for $Q_\lambda(\hat{\boldsymbol{\theta}})$ to achieve the minimum value. Because the augmented component in lasso SEM involves absolute values, the conventional gradient or Newton-type methods for optimization (e.g., Chong & Żak, 2013) do not apply. In the R package lslx (latent structure learning extended), Huang (2020) combined coordinate decent and quasi-Newton algorithms to minimize $Q_\lambda(\boldsymbol{\theta})$ (see also G.-X. Yuan et al., 2012). According to Huang (2020), compared to other R packages of SEM with penalized likelihood, lslx can reliably find the minimizer of $Q_\lambda(\boldsymbol{\theta})$ with the least time. In operation, one can select an optimum λ using BIC or a cross-validation criterion. Following the recommendation of Huang (2020), we will use BIC to select λ in using lslx for the analysis in the following sections. The criterion BIC was also used by K. M. Marcoulides and Falk (2018) in comparing the results following Tabu specification searches.

While the formulation of $F_{ml}(\boldsymbol{\theta})$ is the same across LM test and lasso SEM, the mechanisms behind the two methods are different. LM test releases the constraint $v_j = 0$ according to whether $\hat{\lambda}_j$ is statistically significant, which is asymptotically equivalent to whether the corresponding $\hat{v}_j$ is statistically significant in a model with v_j being freely estimated (Buse, 1982). In contrast, lasso SEM aims to release constraint $v_j = 0$ according to the absolute value of $\hat{v}_j$ in the estimation process, although the component $F_{ml}(\boldsymbol{\theta})$ within the augmented function $Q_\lambda(\boldsymbol{\theta})$ still accounts for the variances and covariances of the parameter estimates. This motivates us to study to what degree the performance of lasso SEM depends on the value of $|v_j|$ and on the effect size $\delta_j = |v_j|/\mathrm{sd}_j$, where sd_j is the square root of the diagonal element of the inverse of the information matrix corresponding to $F_{ml}(\boldsymbol{\theta})$ and evaluated at the population values of $\boldsymbol{\theta}$.

In summary, the elements of $\boldsymbol{\lambda}$ in LM test are subject to estimation, and the significance of each individual $\hat{\lambda}_j$ is also affected by the distribution of the data. In lasso SEM, the value of λ is subject to choice. Thus, although λ is a scalar in lasso SEM, one has to minimize $Q_\lambda(\boldsymbol{\theta})$ across many values of λ, and then chooses one according to a model selection criterion (e.g., AIC, BIC or cross-validation).

Data

In this section, we will first describe a real data set and the corresponding model before introducing two transformations. The transformed data sets play the role of the populations, and the samples of our study are from such populations. The idea of using real data sets to validate a statistical method was used in MacCallum et al. (1992). Our study is different from that of MacCallum et al. in the following aspects: 1) We compare LM test against lasso SEM rather than studying the validity of the LM test itself. 2) While we also rely on a real data set for the comparison to be realistic, the raw data set is transformed so that the "population covariance" matrix in our study is known and so are the locations and population values of the parameters in a correctly specified model. 3) Instead of examining the performance of each method using fit indices or cross-validation index, we use the frequency of selecting the right parameters as the indicator for good performance. 4) We also examine the relationship of frequencies being selected with the absolute values of the non-zero parameters as well as with the effect sizes of the parameters in the population. 5) We use robust transformation to deal with nonnormally distributed data.

Our data come from the website, https://openpsychometrics.org/_rawdata/, and are free to download. The data set IPIP240 (International Personality Item Pool) consists of 240 personality items (see e.g., https://ipip.ori.org/newItemTranslations.htm for IPIP in different languages), designed to measure 6 traits (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, Openness to Experience). All the items are on a 7-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neutral, 5 = slightly agree, 6 = agree, 7 = strongly agree). Each of the 6 traits contains 4 facets, and each facet contains 10 items. Details on the wording of the 240 items can be found at https://ipip.ori.org/newhexaco_pi_key.htm. Twenty three of the 24 facets consist of both positively and negatively worded items. Within each facet, we separately computed the average of the scores on negatively worded items and that of the scores on positively worded ones, and obtained $p = 47$ manifest variables of facet scores for each participant. The data set on the web has 22786 cases, with participants from many countries. In addition to the 240 personality items, two additional items in the questionnaire are "I understand the instructions for this test." and "I have answered all of these questions as accurately as possible." Participants were instructed to also endorse these two items using a 7-point Likert scale. By restricting to USA participants who endorsed the two additional items both at 7, we obtained a sample of $N = 8884$ and $p = 47$. Our "population distribution" will be obtained by further transforming this sample.

We need to have a model in order to examine the performances of the LM test in model modification and lasso SEM in choosing model parameters. By design, the 47 manifest variables should load on six factors in a confirmatory factor model, without cross-loadings. However, fitting the factor model to the sample using NML results in $T_{ml} = 124733.365$, more than 122 times of the nominal degrees of freedom $df = 1019$. Thus, the unidimensional factor model is far from being adequate in describing the relationship of the 47 variables. We then used exploratory factor analysis and NML estimation to identify a better model, while the number of common factor remains to be 6. With the option of promax rotation (kappa = 5) in SPSS and by omitting loadings less than .20, we obtained a pattern matrix with 94 loadings. The SPSS output can be downloaded at https://www3.nd.edu/~kyuan/lassoSEM/SPSS_p47m6.pdf. While most of the variables mainly loaded on the factor they are designed to measure, the loadings of V5 (Honesty-Humility /Greed Avoidance facet, positively keyed) and V7 (Honesty-Humility/Modesty facet, positively keyed) on the first factor (Honesty-Humility) are .180 and .049, respectively. Following the substantive theory from the design, we also add loadings $\gamma_{5,1}$ and $\gamma_{7,1}$ in the model, which has a total of 96 factor loadings, 15 factor correlations, and 47 error variances. The following NML estimation of the confirmatory factor model with 158 free parameters yields $\hat{\gamma}_{5,1} = .374$ with a z-statistic of 23.419, and $\hat{\gamma}_{7,1} = .214$ with a z-statistic of 11.820, validating the substantive theory. However, the corresponding statistic, $T_{ml} = 84849.213$, is still highly significant when referred to the nominal χ^2_{970}. Thus, the 158-parameter model is still far from being adequate in accounting for the covariances of the 47 variables.

To create an empirical distribution whose covariance structure is known, we make the following transformation

$$\mathbf{x}_{i0} = \mathbf{\Sigma}_p^{1/2}\mathbf{S}^{-1/2}\mathbf{x}_i, \quad i = 1, 2, \ldots, 8884, \tag{6}$$

where $\mathbf{S}$ is the sample covariance matrix of the 47 manifest variables, $\mathbf{\Sigma}_p = \mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix with $\hat{\boldsymbol{\theta}}$ being the NML estimates of the 158 parameters, and the superscript 1/2 is power[1] for matrices. Because both $\mathbf{\Sigma}_p^{1/2}$ and $\mathbf{S}^{-1/2}$ are symmetric by definition, standard covariance algebra (e.g., https://en.wikipedia.org/wiki/Covariance) implies that the sample covariance matrix of the x_{i0} is given by

$$\mathbf{S}_0 = \mathbf{\Sigma}_p^{1/2}\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2}\mathbf{\Sigma}_p^{1/2} = \mathbf{\Sigma}_p^{1/2}\mathbf{\Sigma}_p^{1/2} = \mathbf{\Sigma}_p.$$

Thus, the 158-parameter model is a correct covariance structure model for the 47-dimensional random vector $\mathbf{x}_0$ uniformly distributed over the data points $\mathbf{x}_{i0}$, i.e.,

$$P(\mathbf{x}_0 = \mathbf{x}_{i0}) = 1/8884, \quad i = 1, 2, \ldots, 8884. \tag{7}$$

The values of the 158 parameters for the population $\mathbf{x}_0$ are given by $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$. We will refer the transformation in equation (6) as the H_0-transformation. The $\mathbf{x}_{i0}$ in equation (6) or the $\mathbf{x}_0$ in equation (7) serves as one of the populations for us to compare LM test and lasso SEM. We will draw samples of different sizes from $\mathbf{x}_0$ (i.e., from the $\mathbf{x}_{i0}$ with replacement as in bootstrap resampling) and evaluate the performances of the two methods on these samples. While we know the population value of each of the 158 parameters, we do not have a parametric form for the distribution of the 47 variables. The uniform distribution in equation (7) is simply the empirical distribution of the $\mathbf{x}_{i0}$. Such a distribution allows us to rigorously evaluate the performances of the two methods with real data.

Note that the population $\mathbf{x}_0$ represented by the $\mathbf{x}_{i0}$ in equation (6) is nonnormally distributed according to the measure of multivariate kurtosis (Mardia, 1970). Actually, the sample multivariate kurtosis of the $\mathbf{x}_{i0}$ in equation (6) is the same as that of the $\mathbf{x}_i$ in the same equation, and its standardized value is $\text{Mkurt}_s = 198.119$, highly significant when referred to $N(0, 1)$. Thus, a robust method is preferred to account for the distribution property of the $\mathbf{x}_i$. We choose M-estimation for such a purpose. The idea of M-estimation is to weight individual observations differently in estimating the population means and covariance matrix, and cases that are farther away from the center get smaller weights than those located near the center of the distribution of the sample. In particular, the weight for each case can be obtained via a function of the squared Mahalanobis distance d_i^2, and we will use Huber-weight function in our analysis. The Huber-weight function contains a tuning parameter φ that determines a cutoff value via the quantile of the chi-square distribution χ^2_p (see e.g., Yuan & Zhong, 2013). The NML method corresponds to $\varphi = 0$ in which no cases are downweighted, or the cutoff value is at infinity. When setting $\varphi = .05$, the cutoff value is the 95th quantile of χ^2_p, and about 5% of the cases are downweighted if the population[2] is normally distributed. With a larger φ, more cases are downweighted and the extent of the downweighting on these cases also becomes more severe. After the tuning parameter φ is chosen, the robust estimates of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ are computed iteratively, and the weights w_{i1} for computing the mean vector and w_{i2} for computing the covariance matrix are also updated accordingly. At the convergence, these are given by

$$\hat{\boldsymbol{\mu}}_r = \frac{\sum_{i=1}^N w_{i1}\mathbf{x}_i}{\sum_{i=1}^N w_{i1}}, \quad \text{and} \quad \hat{\mathbf{\Sigma}}_r = \frac{1}{N}\sum_{i=1}^N w_{i2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)'.$$

For a proper weighting scheme, the robust estimates $\hat{\boldsymbol{\mu}}_r$ and $\hat{\mathbf{\Sigma}}_r$ are close to most efficient. By noticing that $\hat{\mathbf{\Sigma}}_r$ is the sample covariance matrix of

$$\mathbf{x}_{ir} = w_{i2}^{1/2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r) + \hat{\boldsymbol{\mu}}_r, \tag{8}$$

Yuan et al. (2000) proposed to use (8) as a robust transformation technique. Based on the logic that the sample covariance matrix is asymptotically most efficient when data are normally distributed, Yuan et al. (2000) further proposed choosing the tuning parameter in Huber-weight function according to the sample multivariate kurtosis. In particular, a properly chosen tuning parameter φ should yield a sample $\mathbf{x}_{ir}$ whose standardized multivariate kurtosis is not statistically

[1]For a $p \times p$ non-negative definite matrix $\mathbf{A}$, its power can be obtained by $\mathbf{A}^c = \mathbf{V}\mathbf{B}^c\mathbf{V}'$, where $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$ with $\mathbf{v}_j$ being the jth eigenvector of $\mathbf{A}$ corresponding to the jth eigenvalue b_j, and $\mathbf{B}^c$ is the diagonal matrix with jth diagonal element being given by b_j to the power of c.

[2]For a sample from a normally distributed population, the squared Mahalanobis distance d_i^2 approximately follow χ^2_p.

significant. In this article, we further propose to also compare LM test and lasso SEM over the robustly transformed sample to see how each is affected by nonnormality in practice. In particular, after replacing the $\mathbf{x}_i$ and $\mathbf{S}$ in equation (6), respectively, by the $\mathbf{x}_{ir}$ and $\hat{\mathbf{\Sigma}}_r$, we get a further transformed sample $\mathbf{x}_{ir0}$ whose covariance matrix remains to be $\mathbf{\Sigma}_p = \mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$. Thus, the 158-parameter model is also a correct model for the population $\mathbf{x}_{r0}$ defined by $P(\mathbf{x}_{r0} = \mathbf{x}_{ir0}) = 1/8884$, $i = 1, 2, \ldots, 8884$. We will refer to the combination of the robust transformation and the H_0 transformation as the RH_0-transformation.

Following the suggestion of Yuan et al. (2000), we applied Huber-weight with $\varphi = .05$ to the data set $\mathbf{x}_i$, $i = 1, 2, \ldots, 8884$, and the standardized multivariate kurtosis of the resulting $\mathbf{x}_{ir}$ is $\text{Mkurt}_s = 63.094$, still highly significant. We increased the tuning parameter φ by .05 each time, obtained $\text{Mkurt}_s = 2.436$ at $\varphi = .25$, and $\text{Mkurt}_s = -6.161$ at $\varphi = .30$. Further fine-tuning yields $\text{Mkurt}_s = -1.203$ at $\varphi = .27$, and the corresponding $\mathbf{x}_{r0}$ represented by the $\mathbf{x}_{ir0}$ following the RH_0-transformation serves as the other population for us to compare LM test and lasso SEM. Samples of different sizes will be drawn from the $\mathbf{x}_{r0}$ and are used to compare the performances of the two methods.

Note that the adequacy of a sample size is best judged with the number of variables, and vice versa (see e.g., Yuan et al., 2019). Because the advantage of lasso regularization is for conditions with many variables and relatively small N, we choose $N = 80$, 100, 120, 150, 200 and 300. With $p = 47$, the values of the ratio N/p are 1.70, 2.13, 2.55, 3.19, 4.26, 6.38, respectively. These sample sizes are considered inadequate for the LM test to yield reliable results (MacCallum et al., 1992), and we expect lasso SEM to have an edge. The design allows us to evaluate whether lasso SEM can truly deliver the expected promises in practice. The performances of the two methods under the two populations $\mathbf{x}_0$ and $\mathbf{x}_{r0}$ also allow us to evaluate to what degree each is affected by nonnormality.

At each sample size, 5 independent replications are drawn, giving a total of 30 data sets from the populations $\mathbf{x}_0$ and $\mathbf{x}_{r0}$, respectively. In a typical Monte Carlo study, such a number of replications is considered inadequate. A small number of replications is used in our study because it is time-consuming to administer model modifications following either lasso SEM or LM test. In particular, we expect LM test to identify the 49 cross-loadings that are not in the initial theoretical model. This needs to be achieved by 49 consecutive model modifications and re-estimations, and each is carried out manually. As to be described in the following section, in executing lasso SEM, it is also time-consuming to tune the Lagrangian multiplier λ.

Because it is time-consuming to conduct model modifications, MacCallum et al. (1992) used 10 replications and 4 model modifications for each replication in their study 1, and 3 replications with 3 model modifications for each replication in their study 2. They had a total of 356 model modifications. As to be reported, we have a total of 3084 model modifications following the LM test.

Table 1 summarizes the information on the raw data, the transformed data that play the role of the population, the sample size, and the number of replications. Lasso SEM and LM test will be evaluated under these conditions in the following sections.

Table 1. Data and Conditions.

Raw data	47 facet scores of the IPIP instrument with 8884 individuals
Transformation	H_0: a population with a known covariance structure but heavy tails RH_0: a population with a known covariance structure and non-significant multivariate kurtosis
Sample size	80, 100, 120, 150, 200, 300
# of Replication	5

Method

As noted earlier, the LM test in our study is carried out using EQS 6.4 (Bentler, 2006), and lasso SEM is carried out using lslx 0.6.9 (Huang, 2020). In this section we describe how the two methods are implemented via these packages.

LM test

We start with the unidimensional 6-factor model with 47 free factor loadings, each of the factor variances is fixed at 1.0, all the factor correlations and error variances are free parameters. All the error covariances are fixed at 0. Let's name this model M_0, which is estimated using NML. There are $h = 47 \times 5 = 235$ factor loadings in M_0 that are fixed at 0. Among these 235 loadings, 49 of them are not zero in the population, and the other 186 parameters are truly zero in the population. The LM test in EQS considers the $h = 235$ constraints, and each corresponding to a Lagrange multiplier λ_j. The method of LM test estimates each of the λ_j by solving equation (4) and obtains its standard error (SE) using the augmented likelihood function via equation (1) (see e.g., Bentler & Dijkstra, 1985). The z-score for each of the $\hat{\lambda}_j$ is obtained, and the squared value of the z-score is the expected drop of the likelihood ratio statistic T_{ml} (due to the approximate equivalence between the LM test and the likelihood ratio test) if the parameter becomes freely estimated. The default output of EQS gives the ordered values of the squared z-score for $\hat{\lambda}_j$.

We choose the loading corresponding to the largest drop of T_{ml} to be included in the model, and name it M_1, which is again estimated using NML. The parameter corresponding to the largest drop of T_{ml} following the estimation of M_1 is again selected into the model, and name it M_2, which is again estimated by NML. We continue this process for each sample until LM test[3] does not suggest any additional factor loadings to include, and the last model is denoted as $M_{q_{LM}}$, which has q_{LM} additional factor loadings than that of M_0. For the q_{LM} loadings sequentially identified by LM test, some correspond to the 49 non-zero loadings in the population but not in M_0, and they are correctly identified. Whereas the others correspond to the loadings that are fixed at zero in the population, and they are wrongly identified. The frequency of the correctly identified loadings across replications reflects the power of the LM test, while the frequency of the wrongly identified

[3]The LM test in EQS uses.05 as the default significance level, and the procedure stops when the p-values corresponding to all the $\hat{\lambda}_j$ are greater than.05.

Table 2. Times and frequency that LM and lasso select the non-zero factor loadings in the population: N_t is the total number of loadings selected across 5 replications at a given sample size N; N_c is the times true non-zero factor loadings are selected; $r_c = N_c/N_t$; the Sum line is the summation of the previous six lines, and the value of r_c on the Sum line is the ratio of the summed N_c over the summed N_t; the confidence interval (CI) is for the population probability of correctly selecting the non-zero factor loadings.

	LM(H_0)			Lasso(H_0)			LM(RH_0)			Lasso(RH_0)		
N	N_t	N_c	r_c	N_t	N_c	r_c	N_t	N_c	r_c	N_t	N_c	r_c
80	227	136	.599	232	142	.612	202	139	.688	256	156	.609
100	244	164	.672	338	175	.518	226	151	.668	364	186	.511
120	256	169	.660	385	209	.543	273	192	.703	386	212	.549
150	265	201	.758	391	218	.558	266	201	.756	386	218	.565
200	270	205	.759	413	229	.554	268	217	.810	393	230	.585
300	306	226	.739	439	238	.542	282	235	.833	454	244	.537
Sum	1568	1101	.702	2198	1211	.551	1517	1135	.748	2239	1246	.556
CI		[.672,.732]			[.531,.571]			[.716,.781]			[.540,.573]	

Note: H_0 represents the population $\mathbf{x}_0$ defined by the $\mathbf{x}_{i0}$ in equation (6), and RH_0 represents the population x_{r0} defined by equations (8) and (6).

ones reflects the type I error rate of the LM test, due to capitalization on chance. For each condition of N, and under the transformations H_0 and RH_0, the total number of identified loadings N_t (i.e., the sum of the q_{LM} across the 5 replications), the number of correctly identified loadings (N_c), as well as the ratio of correct identification across the 5 replications ($r_c = N_c/N_t$) are recorded, and are reported in Table 2.

Lasso SEM

The carrying out of lasso SEM also starts with the unidimensional model M_0, with $h = 5 \times 47 = 235$ factor loadings in the penalized term. The key element of lasso SEM is the tuning parameter λ, which controls the contribution of $g(\lambda, \boldsymbol{\theta})$ in equation (5) to the augmented function $Q_\lambda(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \lambda)$ in equation (1). In using the R package lslx (Huang, 2020), we start with tuning the range of the Lagrange multiplier λ. At $\lambda = 1.50$, no loading is selected from the h candidates by lslx for the 60 samples (5 replications by 6 sample sizes, and under the H_0- and RH_0-transformations). We thus choose the interval [0, 1.5] as the range of λ. Within this range, we choose 151 values of λ: 0, .01, .02, …, 1.49, 1.50. For each of the 60 sample covariance matrices, and at each chosen λ, the function $Q_\lambda(\boldsymbol{\theta})$ is minimized using lslx. There are a few samples that the R package gives a warning that the algorithm for minimizing the function $Q_\lambda(\boldsymbol{\theta})$ did not converge, and these results are discarded. For each sample, the set of non-zero loadings corresponding to the λ with the smallest BIC is recorded,[4] and the resulting model is regarded as the best model endorsed by lasso SEM. Parallel to the results under LM test, at each sample size, the total number (N_t) of parameters being selected by lasso SEM across the 5 replications is recorded, N_c of the N_t parameters are from the set of the 49 non-zero loadings that were not included in M_0, while the other ($N_t - N_c$) parameters correspond to the loadings whose values are zero in the population. The frequency of the correctly identified non-zero loadings across replications reflects the power of the lasso methodology, and the frequency of the wrongly identified ones reflects the type I error rate. The values of N_t, N_c and $r_c = N_c/N_t$ are presented under the transformations H_0 and RH_0 in Table 2, respectively.

Evaluation criteria

We used $r_c = N_c/N_t$ to compute the rate of correctly identifying the non-zero parameters, and $r_w = 1 - r_c$ is the rate of wrongly identifying the parameters whose population values are zero. Thus, the value of r_c reflects the power properties of the method, whereas the value of r_w can be regarded as the type I error rate. Because $r_c + r_w = 1$, a method with better power also has a better control of type I errors. Note that, unlike in a conventional statistical test, the perfect type I error rate under this setup is 0 instead of .05.

Another approach to evaluate the performances of the two methods is to use $r_c^* = N_c/q_c$ and $r_w^* = (N_t - N_c)/q_w$, where $q_c = 5 \times 49 = 245$ is the maximum number of selecting non-zero loadings in the population (correct selection) across the 5 replications, and $q_w = 5(5 \times 47 - 49) = 930$ is the maximum number of selecting zero loadings in the population (wrong selection) across the 5 replications. Correct selections and wrong selections are weighted differently under this approach. Such an approach might be proper in classifications where competing methods are forced to select equal number of individuals. It may not be proper here because lasso SEM tends to select many more parameters than LM test does. Also, it becomes hard to compare lasso SEM against LM test or any other method because $r_c^* + r_w^* \neq 1.0$, and a method may have both values of r_c^* and r_w^* close to 1.0.

We noted earlier that lasso SEM selects parameters according to absolute values by equation (5). In contrast, LM test is a pure statistical method and we may expect its rate of correct selection to depend more on the effect sizes corresponding to the non-zero loadings. To see how the rate of a correct selection is related to the absolute value of the non-zero loading v_j as well as to the effect size $\delta_j = v_j/sd_j$, the times for selecting each of the 49 non-zero loadings across the 6 conditions of sample size and 5 replications are combined, and let this number be N_{cj}, $j = 1, 2, \ldots, 49$. Thus, the value of N_{cj} for a perfect selection is 30. But N_{cj} can also be zero if the method performs poorly. We will plot the value N_{cj} against the values of $|v_j|$ and δ_j, respectively. We also use the tool of plot to compare the values of N_{cj} between different methods and under different transformations. Such graphics together with least-squares regression lines and the associated values of R^2 allow us

[4]The package lslx has two rules of indicating non-zero parameters or parameters being selected form the penalized list. One is "Coefficient test" and the other is "Numerical conditions". The two rules do not always give the same results, and we used the results of "Numerical conditions".

to answer the question as to what extent the rate of correct selection is related to the value of $|v_j|$ and δ_j. They also allow us to see to what degree each method is affected by the nonnormality of the population distribution as measured by Mardia's multivariate kurtosis.

Results and Analysis

With 49 non-zero cross-loadings and 5 replications, we would expect an ideal method to select $5 \times 49 = 245$ additional parameters at each condition of N. The results in Table 2 indicate that both LM test and lasso SEM select more parameters as the sample size N increases. For samples under H_0 and RH_0, LM test selects few parameters than expected at $N = 80$ and 100, and more parameters than expected when $N \geq 120$. Lasso SEM selects few parameters than expected only at $N = 80$ under H_0, and more parameters than expected in all the other conditions. The results indicate that, for the total number of parameters selected (N_t), lasso SEM is rather sensitive to N.

As N increases, the ratio of correct selection (r_c) under LM (H_0) tends to increase but with fluctuations. Under LM(RH_0), the pattern for r_c to increase with N is more clear. For the 6 conditions of N, four of the values of r_c under LM(RH_0) are greater than their counterparts under LM(H_0), indicating that LM test performs more reliably under the robust transformation.

While the number of selected parameters (N_t) by lasso SEM increases with sample size, the values of r_c do not show a pattern for the statistical power to increase with N. Also, lasso SEM is insensitive to nonnormality nor to the robust transformation.

Under H_0, for the 6 conditions of N, only at $N = 80$ the ratio r_c in selecting the true non-zero loadings by the LM test is smaller than that by lasso SEM. For the other 5 conditions of N, the r_cs following the LM test are greater than those following lasso SEM. Under RH_0, the values of r_c by the LM test are uniformly greater than those by lasso SEM.

The Sum line of Table 2 contains the summation of the values of N_t and N_c for the six conditions of N, respectively. The value of r_c on this line is computed as the ratio of the summed N_c over the summed N_t, and it reflects the overall power for each method. The summed r_c under LM(RH_0) is .748, greater than its counterpart of .702 under LM(H_0). In contrast, the summed r_c by lasso SEM are .551 under H_0 and .556 under RH_0. They are below their LM test counterparts by .151 and .192, respectively. Note that $r_w = 1 - r_c$, and a greater value of correct selection (power) in Table 2 also implies a smaller value of wrong selection (type I errors). Thus, lasso SEM has less power in selecting true non-zero factor loadings and greater type I errors in selecting loadings that are truly 0 in the population.

The last line of Table 2 contains the 95% confidence interval (CI) for the power of each procedure to correctly identify the non-zero loadings in the population. The intervals are computed by $[r_c - 1.96\text{SE}, r_c + 1.96\text{SE}]$, where the r_c is the overall power rate reported in the Sum line and the SE is the standard error computed using the formula provided in the appendix of this article. According to the results, the CI under LM(H_0) overlaps with that under LM(RH_0). But the two CIs under LM test are well separated from those under lasso SEM, indicating the former is more effective in selecting true non-zero factor loadings.

Note that the number of non-zero loadings selected by lasso SEM is uniformly greater than that by LM test, because lasso SEM always select more parameters, especially at larger values of N. Thus, if we do not care about the total number of parameters selected, then lasso SEM performs "better" than LM test. But such an argument might violate the basic principle of statistical inference. That is, we need to control type I errors in order to compare statistical power. Also, the average numbers of selected parameters across the 6 conditions of N by all the methods are above the ideal number 245. The one closest to the ideal number is 252.7 with LM(RH_0), and followed by 261.3 with LM(H_0), 366.3 with Lasso(H_0), and 373.2 with Lasso(RH_0).

Table 3 contains the values of the 49 non-zero loadings that are not included in the initial model M_0, where the subscripts in $v_{j,k}$ refer to variables and factors, respectively.

Table 3. Values of the 49 non-zero factor loadings $v_{j,k}$ in the population but are excluded from the initial model M_0, effect sizes $\delta_{j,k} = v_{j,k}/\text{sd}_{j,k}$ and the times being selected by LM test and lasso SEM across the 6 conditions of N and 5 replications.

$v_{j,k}$	loading	ef-size	LM(H_0)	Lasso(H_0)	LM(RH_0)	Lasso(RH_0)
$v_{3,2}$	0.514	0.407	29	27	29	28
$v_{3,5}$	0.387	0.319	28	26	27	27
$v_{4,2}$	0.384	0.365	25	26	26	26
$v_{5,3}$	−0.603	0.401	25	24	26	25
$v_{6,3}$	−0.293	0.321	23	19	20	19
$v_{7,2}$	0.384	0.243	17	21	19	23
$v_{7,3}$	−0.266	0.153_2	13_4	16_5	14_4	16_5
$v_{7,4}$	0.441	0.252	16_5	8_1	21	11_2
$v_{7,6}$	−0.478	0.276	27	25	27	26
$v_{8,3}$	−0.585	0.621	27	30	28	30
$v_{9,6}$	−0.565	0.461	29	29	29	29
$v_{10,1}$	0.272	0.192_5	10_3	13_4	12_2	15_4
$v_{10,4}$	-0.233_4	0.167_3	7_2	10_2	13_3	10_1
$v_{10,6}$	−0.608	0.464	30	28	30	29
$v_{12,4}$	−0.298	0.319	23	30	24	30
$v_{13,1}$	-0.091_1	0.108_1	6_1	18	7_1	21
$v_{13,3}$	0.456	0.478	29	29	29	29
$v_{13,6}$	−0.354	0.382	29	27	29	27
$v_{14,1}$	0.384	0.309	17	22	17	22
$v_{14,4}$	0.429	0.347	30	30	29	30
$v_{15,1}$	0.554	0.435	22	23	22	25
$v_{15,3}$	0.418	0.344	28	30	28	30
$v_{17,4}$	−0.454	0.415	28	30	28	30
$v_{19,4}$	−0.289	0.300	22	30	25	29
$v_{20,2}$	0.341	0.330	21	27	23	28
$v_{20,4}$	0.255	0.242	17	22	19	23
$v_{21,2}$	0.254_5	0.237	19	27	23	28
$v_{21,6}$	−0.485	0.411	29	29	29	30
$v_{22,4}$	0.300	0.313	22	28	25	29
$v_{23,2}$	−0.716	0.549	30	30	30	30
$v_{24,2}$	0.302	0.263	25	28	24	29
$v_{25,1}$	0.618	0.505	17	27	19	27
$v_{27,1}$	0.646	0.628	20	28	21	28
$v_{29,1}$	0.515	0.571	17	28	18	28
$v_{30,2}$	−0.269	0.216	17	20	20	22
$v_{31,1}$	0.297	0.228	21	11_3	17	11_2
$v_{31,2}$	−0.508	0.409	26	28	26	29
$v_{35,1}$	0.472	0.434	24	27	26	28
$v_{36,1}$	−0.356	0.420	27	30	28	30
$v_{37,4}$	−0.298	0.254	23	25	22	26
$v_{39,1}$	0.388	0.341	21	26	23	26
$v_{39,2}$	−0.278	0.248	20	25	20	26
$v_{40,2}$	0.369	0.434	27	29	28	30
$v_{40,4}$	0.278	0.335	21	28	24	28
$v_{41,1}$	0.381	0.335	25	25	25	27
$v_{41,2}$	0.447	0.394	30	29	27	30
$v_{43,1}$	0.179_2	0.170_4	17	16	16_5	17
$v_{47,1}$	0.230_3	0.211	19	20	18	23
$v_{47,5}$	−0.294	0.266	26	27	25	26
Total			1101	1211	1135	1246

The five smallest absolute values in each column are marked in subscript.

The value of the effect size $\delta_{j,k}$ as well as the times (N_c) of selecting $v_{j,k}$ across the 6 conditions of N and 5 replications are also reported in Table 3. The five smallest values in each column are marked using subscripts. The results in the table suggest that both LM test and lasso SEM are affected more by the effect sizes of the parameters than by their absolute values. For example, the effect size for $v_{7,3}$ is the second smallest but the value of $|v_{7,3}|$ is not among the five smallest. Correct selections for this parameter are among the lowest by each method. In contrast, correct selections for $v_{47,1}$ are not among the five lowest although its value (not effect size) is the third smallest.

Figure 1 contains the plots of the times for LM test and lasso SEM to select the 49 non-zero loadings against the absolute values of the loadings. The least-squares (LS) regression line as well as the corresponding R^2 are also added to each plot. The plots show that LM test is more sensitive to the absolute value of the loadings than lasso SEM. However, the values of R^2 for both methods are rather small although they are in the "large" category according to Cohen (1992).

Figure 2 contains the plot of the times for LM test and lasso SEM to correctly select the non-zero loadings against the effect sizes of the loadings. The results indicate that both the methods are more sensitive to the effect sizes than to the absolute values of the loadings. But all the values of the R^2s are still below .50.

In both Figures 1 and 2, the R^2 under LM test is greater after robust transformation, indicating that LM test performs better with data closer to normally distributed. In contrast, the R^2 under lasso SEM becomes smaller with the robustly transformed data, which seems odd. We suspect that this is due to the fact that the augmented component in equation (5) ignores the sampling errors in the estimates $\hat{v}_j$ when counting their contributions to

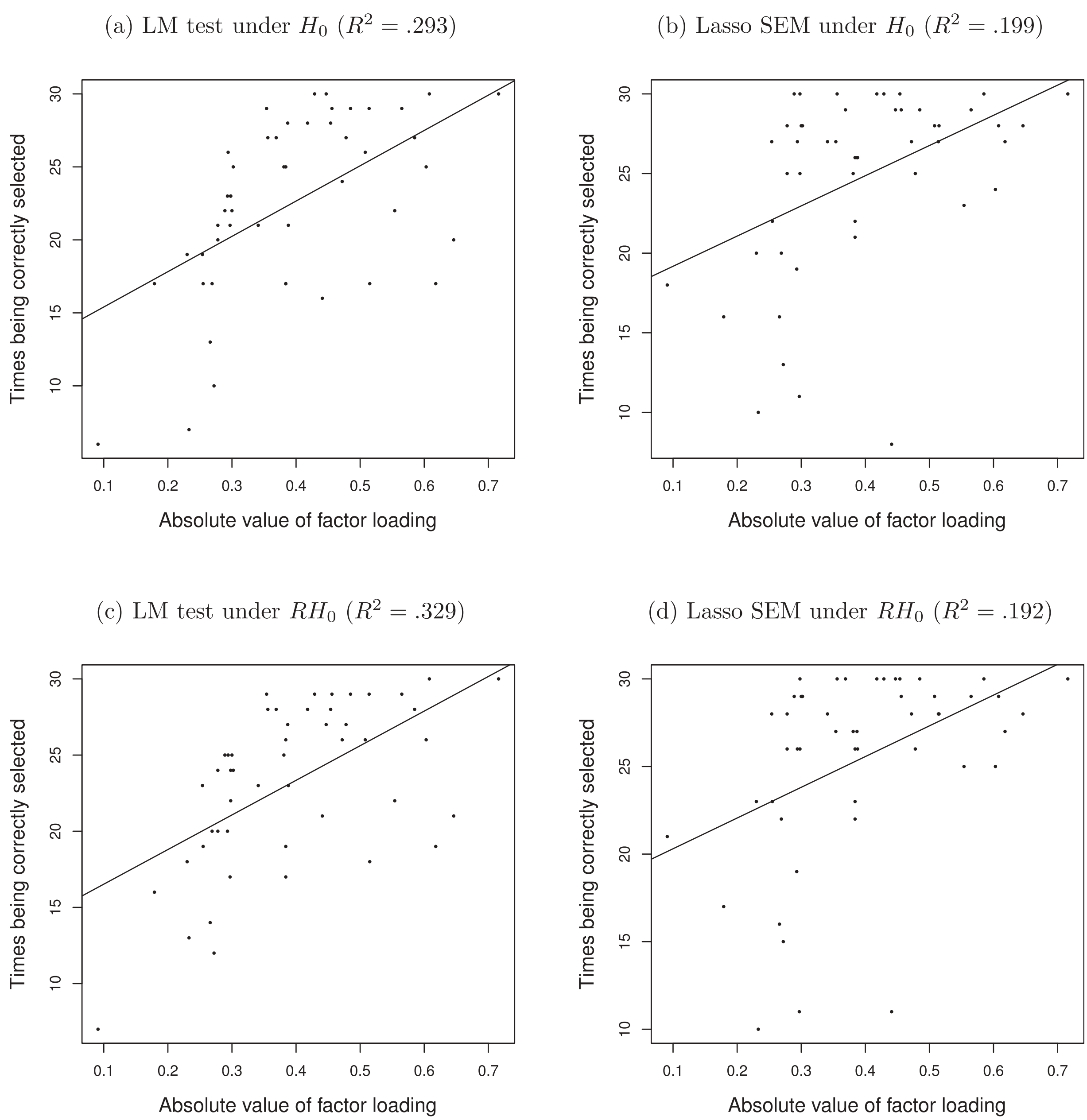

Figure 1. Times for correctly selecting non-zero loadings by LM test and lasso SEM against the size of the factor loading (Table 3).

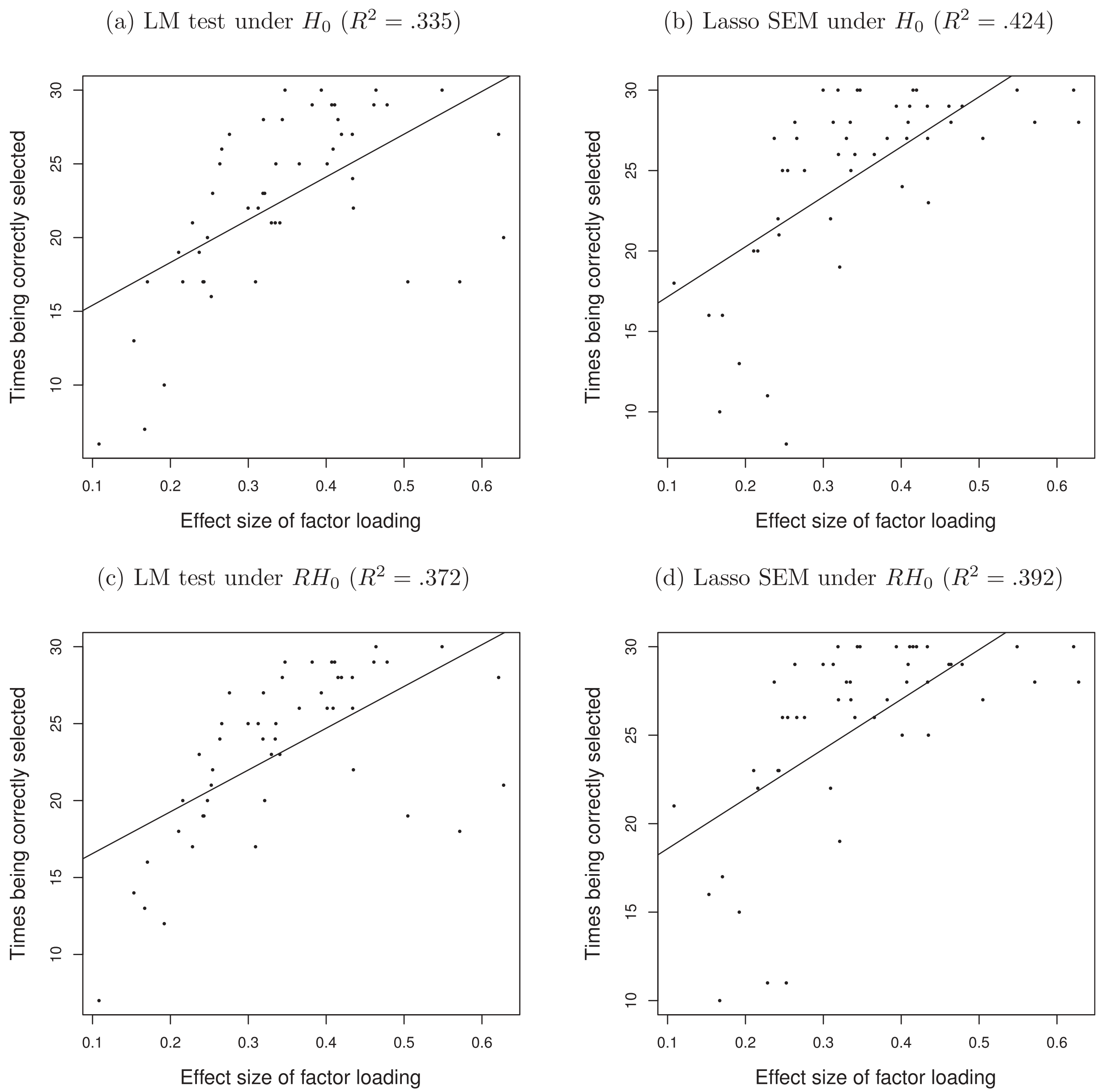

Figure 2. Times for correctly selecting non-zero loadings by LM test and lasso SEM against the effect size of the factor loading (Table 3).

the objective function $Q_\lambda(\boldsymbol{\theta})$ (see equation 1). However, these estimates also appear in the normal-distribution-based $F_{ml}(\boldsymbol{\theta})$, and we expect the method to perform better in all aspects when data are closer to normally distributed. Further research is needed to better understand this issue.

The plots in Figure 3 contrast the numbers of correct selections for LM test and lasso SEM under H_0 against those under RH_0, and for those by lasso SEM against those by LM test. The results indicate that lasso SEM is less affected by robust transformation than LM test. The plots also show that the two methods have a lot of differences in selecting each of the 49 parameters, with $R^2 = .530$ and .560 before and after the robust transformation, respectively.

Note that the plots in Figures 1 to 3 used the numbers of correct selection, not the proportions of correct selection. For the numbers of correct selections reported in Table 3, lasso SEM selects a lot more parameters in total than LM test does, as being reported in Table 2. But the patterns of the scatter plots as well as the R^2 in the figures do not depend on the scale of the variables.

Conclusions and discussion

The main purpose of this article is to contrast the performance of lasso SEM against that of LM test. In addition, we also examined how LM test and lasso SEM are affected by nonnormality of the population distribution, and how lasso SEM is affected by sample size. Our aim is to address the 5 questions, as listed in the introduction section. Based on what we have found, our answers to these questions are:

(1) *Can we trust the results of lasso SEM for model modification*? No.
(2) *Does the performance of lasso SEM depend more on the effect size or the absolute value of the parameter*? It

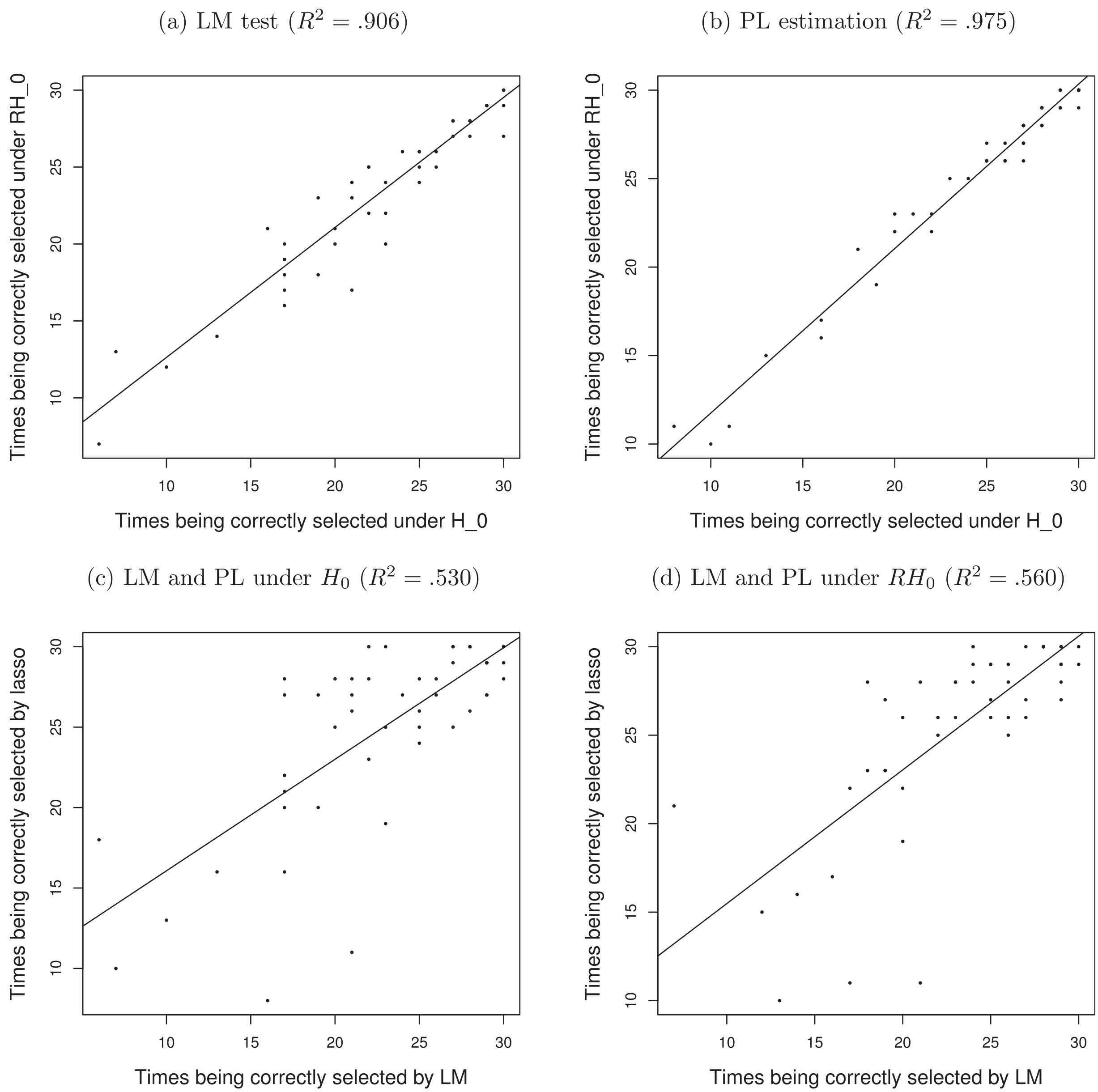

Figure 3. Times for correctly selecting non-zero loadings by LM test and lasso SEM under H_0 and RH_0 (Table 3).

depends more on the effect size than on the absolute value of the parameter.

(3) *Does lasso SEM perform better than the widely used LM test for model modification*? No.

(4) *Are lasso SEM and LM test affected by nonnormally distributed data in practice*? LM test is sensitive to nonnormality. But lasso SEM seems little affected by nonnormality.

(5) *Do lasso SEM and LM test perform better with robustly transformed data*? LM test performs better with robustly transformed data, but not for lasso SEM.

Thus, while LM test is not a reliable procedure to find a model that truly reflects the population relationship among variables, lasso SEM cannot deliver a sound model either. The results of this article indicate that lasso SEM is even less reliable than LM test in locating the truly non-zero parameters in the population.

When N is sufficiently large, the results in MacCallum et al. (1992) suggest that LM test performs well. The results in this article indicate that, in addition to a large N, the robust transformation also increases the reliability of the LM test. In particular, it may not be realistic to increase the sample size in practice. But the robust transformation is a statistical method that can be easily applied, and it is expected to increase the reliability of LM test whenever Mardia's multivariate kurtosis of the observed data is greater than that of the normal distribution. In addition, for LM test to yield more reliable model modification, robust method also yields more accurate parameter estimates as well as more reliable test statistics and fit indices for overall model evaluation (Yuan & Zhong, 2013). Since data in social and behavioral sciences tend to have heavy tails (e.g., Blanca et al., 2013; Cain et al., 2017; Micceri, 1989), we recommend robust methods be routinely applied.

Our results clearly indicate that lasso SEM does not possess the oracle property at the range of the selected N (80–300). According to Huang et al. (2017), the oracle property is an asymptotic property for an adaptive version of lasso. Because there is another tuning parameter with the adaptive lasso (i.e., SCAD and MCP) in addition to the Lagrangian multiplier λ, more guidance is needed on how to choose the tuning parameters. It is also interesting that the performance of lasso SEM does not improve with N (see Table 2). This can be due to the same mechanism behind the phenomenon that lasso SEM is insensitive to the population distribution. Again, we suspect that this is because the augmented component in equation (5) ignores the sampling errors in the estimates $\hat{v}_j$. Given the fact that lasso regularization selects a parameter by the absolute value of its estimate, it is not a surprise that the obtained results might seem against the law of statistics. More studies are needed to rectify such an issue of lasso SEM or that of lasso regularization in general.

In this article, we only compared lasso SEM and LM test for structural equation modeling where a theoretical initial model is available. In practice, when such a model is not available or when the initial model is based on a preliminary hypothesis, exploratory factor analysis, exploratory SEM as well as other heuristic search algorithms can be more effective in delivering a model with a better global fit (e.g., Mai et al., 2018; Marcoulides & Drezner, 2003; Marcoulides et al., 1998; Marcoulides & Leite, 2014). As pointed out by MacCallum et al. (1992, see also K. M. Marcoulides & Falk, 2018), any such data-driven model searching method is susceptible to capitalization on chance. Repeated validations and replications are needed for using the identified models to advance the understanding of subjective matters. Then, model modification will lead to theory modification/advancement.

Our comparison of lasso SEM and LM test is based on their percentages in locating the non-zero parameters in the population. In statistical modeling, freely estimating a parameter whose population value is zero may have a different consequence than excluding a parameter whose population value is non-zero. In particular, excluding a non-zero parameter may cause a systematic bias while including a parameter whose population value is zero may only cause a loss in efficiency. However, researchers in SEM typically rely on the final model to interpret the relationship among the variables, and both types of errors can have categorical substantive consequences. Such an issue can be further discussed given the specific context, which is beyond the scope of this article.

In our study, the criterion BIC built in lslx (Huang, 2020) was used to implement lasso SEM. The results might be different if another criterion was used to select a model from those corresponding to the 151 values of the tuning parameter λ. Based on the findings of MacCallum et al. (1992), it is hard to imagine that another criterion (including cross-validation) will perform substantially better. This is because capitalization on chance is the nature of all statistical methods, which need to reflect what are observed in the current sample. With a relatively small sample size, fluctuations across samples can be substantial. Models obtained from such samples will inherit such properties, and it is hard for them to behave well across replications. Again, repeated validations and replications are needed from model modification to theory advancement.

The poor performance of lasso SEM might be related to the fact that the initial theoretical model in our study is far from being correct. Such a condition also defies the mechanism of LM test and other sequential search methods (Green & Thompson, 2010; K. M. Marcoulides & Falk, 2018). This is because these methods make model changes incrementally and the search algorithm may not traverse the complete model parameter space. When there are many omitted parameters in the current model, an early mistake in the process can interfere with future choices. Essentially the same mechanism also challenges the function of chi-square test statistics in comparing the goodness of nested models when the base model is misspecified (Yuan & Bentler, 2004). In practice, substantial discrepancy between data and model can still exist even when the initial model is formulated based on a substantive theory or according to the design. Then, the automated search procedures as described by K. M. Marcoulides and Falk (2018) might have an edge over an incremental procedure. In particular, K. M. Marcoulides and Falk (2018) showed that the Tabu search procedure can be easily implemented via open-source R functions, and it works well in their illustration with a simulated sample. Additional studies are needed to compare lasso SEM and LM test against these automated search procedures. In such a comparison one might include the multivariate version of the LM test (Bentler, 2006), which is less incremental by evaluating multiple parameters in each step. Also, while the model and population studied in this article are obtained from real data, they may not be general enough. There might be conditions under which lasso SEM outperforms the conventional methods. Future studies should include a variety of model and data types in order to find out such conditions, including models with only a few omitted paths.

Funding

This work was supported by Grant 31971029 from the Natural Science Foundation of China.

ORCID

Ke-Hai Yuan http://orcid.org/0000-0003-0610-1745

References

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis VI*. (pp. 9–42). North-Holland. ISBN: 0444876022.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Multivariate Software. http://www.econ.upf.edu/satorra/CourseSEMVienna2010/EQSManual.pdf

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84. https://doi.org/10.1027/1614-2241/a000057

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician*, *36*, 153–157. https://doi.org/10.1080/00031305.1982.10482817

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence, and estimation. *Behavior Research Methods*, *49*, 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Choi, J., Oehlert, G., & Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, *3*, 429–436. https://doi.org/10.4310/SII.2010.v3.n4.a1

Chong, E. K. P., & Źak, S. H. (2013). *An introduction to optimization* (4th ed.). Wiley.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Fan, J., & Li, R.-Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273

Green, S. B., & Thompson, M. S. (2010). Can specification searches be useful for hypothesis generation? *Journal of Modern Applied Statistical Methods*, *9*, 160–171. https://doi.org/10.22237/jmasm/1272687300

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132. https://doi.org/10.1016/j.csda.2014.05.011

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in a factor analysis model. *Statistics and Computing*, *25*, 863–875. https://doi.org/10.1007/s11222-014-9458-0

Huang, P.-H. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software, 93*, 1–37. https://doi.org/10.18637/jss.v093.i07

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*, 329–354. https://doi.org/10.1007/s11336-017-9566-9

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*, 555–566. https://doi.org/10.1080/10705511.2016.1154793

Lee, S.-Y., & Bentler, P. M. (1980). Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal*, *14*, 121–136. https://www.ingentaconnect.com/content/sabinet/sasj/1980/00000014/00000002/art00009

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mai, Y., Zhang, Z., & Wen, Z. (2018). Comparing exploratory structural equation modeling and existing approaches for multiple regression with latent variables. *Structural Equation Modeling*, *25*, 737–749. https://doi.org/10.1080/10705511.2018.1444993

Marcoulides, G. A., & Leite, W. L. (2014). Exploratory data mining algorithms for conducting searches in structural equation modeling: A comparison of some fit criteria. In J. J. McArdle & G. Ritschard (Eds..), *Contemporary issues in exploratory data mining in the behavioral sciences*. (pp. 150–171). Taylor & Francis. ISBN: 9780415817097.

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling*, *10*, 154–164. https://doi.org/10.1207/S15328007SEM1001_8

Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using tabu search. *Structural Equation Modeling*, *5*, 365–376. https://doi.org/10.1080/10705519809540112

Marcoulides, K. M., & Falk, C. F. (2018). Model Specification Searches in Structural Equation Modeling with R. *Structural Equation Modeling*, *25*, 484–491. https://doi.org/10.1080/10705511.2017.1409074

Mardia, K. V. (1970). Measure of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530. https://doi.org/10.1037/0033-2909.105.1.156

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. https://doi.org/10.1037/00332909.105.1.156

Ning, L., & Georgiou, T. T. (2011). Sparse factor analysis via likelihood and l_1 regularization. In *50th IEEE conference on decision and control and European control conference* (pp. 5188–5192), Orlando, FL. https://doi.org/10.1109/CDC.2011.6161415

Pan, J.-H., Ip, H. S., & Dube, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*, 687–704. https://doi.org/10.1037/met0000112

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*, 50–57. https://doi.org/10.1017/S0305004100023987

Silvey, S. D. (1959). The Lagrangian multiplier test. *Annals of Mathematical Statistics*, *30*, 389–407. https://doi.org/10.1214/aoms/1177706259

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384. https://doi.org/10.1007/BF02294623

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). An improved GLMNET for L1-regularized logistic regression. *Journal of Machine Learning Research*, *13*, 1999–2030. https://doi.org/10.1145/2020408.2020421

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757. https://doi.org/10.1177/0013164404264853

Yuan, K.-H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *53*, 31–50. https://doi.org/10.1348/000711000159169

Yuan, K.-H., Fan, C., & Zhao, Y. (2019). What causes the mean bias of the likelihood ratio statistic with many variables? *Multivariate Behavioral Research*, *54*, 840–855. https://doi.org/10.1080/00273171.2019.1596060

Yuan, K.-H., Marshall, L. L., & Weston, R. (2002). Cross-validation through downweighting influential cases in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *55*, 125–143. https://doi.org/10.1348/000711002159734

Yuan, K.-H., & Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychological Methods*, *18*, 121–136. https://doi.org/10.1037/a0031604

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429. https://doi.org/10.1198/016214506000000735

Appendix

In this appendix, we derive a formula for computing the standard error (SE) of the overall rate of correctly identifying the non-zero loadings in the population by each procedure, as reported in the Sum line of Table 2. The obtained formula of SE is used to compute the confidence interval (CI) for the probability (i.e., power) of correctly selecting the non-zero factor loadings.

Let i be the index for replication, j be the index for the 49 non-zero loadings that were excluded from the initial model M_0, and k for the level of sample size. Let $x_{ij}^{(k)} = 1$ if the jth parameter is identified as non-zero by a method in the ith replication for the kth level of sample size, and $x_{ij}^{(k)} = 0$ otherwise. The value of r_c in the Sum line of Table 2 is obtained according to

$$r_c = \frac{1}{N_s}\sum_{k=1}^{6}\sum_{i=1}^{5}\sum_{j=1}^{49} x_{ij}^{(k)},$$

where N_s is the sum of the total number of parameters being selected across the 5 replications and 6 conditions of sample size.

Note that the replications within a condition of sample size and across different conditions of sample size are independent in our study. Consequently, $x_{ij}^{(k)}$ are independent across different values of i and/or k. However, $x_{ij_1}^{(k)}$ and $x_{ij_2}^{(k)}$ are not independent for given (i,k), simply because they are from the same sample. Thus,

$$\begin{aligned}\sigma_r^2 &= \mathrm{Var}(r_c)\\ &= \tfrac{1}{N_s^2}\sum_{k=1}^{6}\sum_{i=1}^{5}\mathrm{Var}\Big(\sum_{j=1}^{49}x_{ij}^{(k)}\Big)\\ &= \tfrac{1}{N_s^2}\sum_{k=1}^{6}\sum_{i=1}^{5}\Big(\sum_{j=1}^{49}\sigma_{jj}^{(k)}+\sum_{j_1\neq j_2}\sigma_{j_1j_2}^{(k)}\Big)\\ &= \tfrac{5}{N_s^2}\sum_{k=1}^{6}\Big(\sum_{j=1}^{49}\sigma_{jj}^{(k)}+\sum_{j_1\neq j_2}\sigma_{j_1j_2}^{(k)}\Big),\end{aligned} \tag{A1}$$

where $\sigma_{jj}^{(k)}=\mathrm{Var}(x_{ij}^{(k)})$ and $\sigma_{j_1j_2}^{(k)}=\mathrm{Cov}(x_{ij_1}^{(k)},x_{ij_2}^{(k)})$. Let $\bar{x}_j^{(k)}=\sum_{i=1}^{5}x_{ij}^{(k)}/5$. Then

$$\hat{\sigma}_{jj}^{(k)}=\frac{1}{4}\sum_{i=1}^{5}(x_{ij}^{(k)}-\bar{x}_j^{(k)})^2 \quad\text{and}\quad \hat{\sigma}_{j_1j_2}^{(k)}=\frac{1}{4}\sum_{i=1}^{5}(x_{ij_1}^{(k)}-\bar{x}_{j_1}^{(k)})(x_{ij_2}^{(k)}-\bar{x}_{j_2}^{(k)})$$

are unbiased estimates of $\sigma_{jj}^{(k)}$ and $\sigma_{j_1j_2}^{(k)}$, respectively. Substituting these values into equation (A1) yields an unbiased estimate $\hat{\sigma}_r^2$.

According to Table 2, the values of N_s for methods LM(H_0), Lasso(H_0), LM(RH_0), and Lasso(RH_0) are 1568, 2198, 1517, and 2239, respectively. The corresponding standard errors for the four estimates (r_c) 0.702, 0.551, 0.748, and 0.556 are 0.0154, 0.0102, 0.0166, and 0.0086, respectively. The confidence intervals reported in the last line of Table 2 are obtained according to $[r_c-1.96\hat{\sigma}_r, r_c+1.96\hat{\sigma}_r]$.