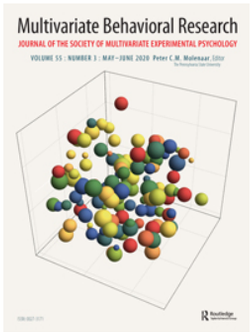

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Postselection Inference in Structural Equation Modeling

Po-Hsien Huang

To cite this article: Po-Hsien Huang (2020) Postselection Inference in Structural Equation Modeling, Multivariate Behavioral Research, 55:3, 344-360, DOI: 10.1080/00273171.2019.1634996

To link to this article: https://doi.org/10.1080/00273171.2019.1634996

Published online: 13 Jul 2019.

MULTIVARIATE BEHAVIORAL RESEARCH
2020, VOL. 55, NO. 3, 344–360
https://doi.org/10.1080/00273171.2019.1634996

Postselection Inference in Structural Equation Modeling

Po-Hsien Huang

Department of Psychology, National Cheng Kung University

ABSTRACT

Most statistical inference methods were established under the assumption that the fitted model is known in advance. In practice, however, researchers often obtain their final model by some data-driven selection process. The selection process makes the finally fitted model random, and it also influences the sampling distribution of the estimator. Therefore, implementing naive inference methods may result in wrong conclusions—which is probably a prime source of the reproducibility crisis in psychological science. The present study accommodates three valid state-of-the-art postselection inference methods for structural equation modeling (SEM) from the statistical literature: data splitting (DS), postselection inference (PoSI), and the polyhedral (PH) method. A simulation is conducted to compare the three methods with the commonly used naive procedure under selection events made by L_1-penalized SEM. The results show that the naive method often yields incorrect inference, and that the valid methods control the coverage rate in most cases with their own pros and cons. Real world data examples show the practical use of the valid inference methods.

KEYWORDS

Structural equation modeling; factor analysis; lasso; postselection inference; polyhedral lemma

Introduction

Model selection is a commonly used strategy for structural equation modeling (SEM). It is often advocated as a confirmatory method for multi-model inference, that is, drawing research conclusions based on several pre-specified candidate models (e.g., Burnham & Anderson, 2002; MacCallum & Austin, 2000). Sometimes, model selection is conducted as an exploratory method for searching a best-fitting model (e.g., Chou & Bentler, 1990; MacCallum, 1986). For example, an SEM user may initialize a tentative model and then empirically improve it step by step according to the statistical significance implied by Lagrange multipliers (Silvey, 1959) or Wald tests (Wald, 1943). The exploratory case is sometimes called model generation (e.g., Jöreskog, 1993).

Standard large sample theory is applicable for statistical inference under the assumption that the considered model is fixed (not random). Regardless of the model selection method, the selection process makes the" fixed model" assumption invalid. This fact was recognized by statisticians about five decades ago (e.g., Brown, 1967; Buehler & Feddersen, 1963). More recently, Leeb and his colleagues demonstrated in a series of works how the presence of model selection destroys the asymptotic normality of the parameter estimator (e.g., Kabaila & Leeb, 2006; Leeb & Pötscher, 2003, 2005, 2006, 2008). Breiman (1992) even declared it" a quiet scandal in the statistical community" to perform statistical inference without model selection. Nevertheless, making a valid postselection inference is a notoriously difficult task (see Leeb & Pötscher, 2006, for some negative results)—that is why the issue of postselection inference is still unresolved.

To demonstrate how model selection may invalidate traditional inference, let us consider the following example originally presented by Benjamini and Yekutieli (2005). Let $\{T_j\}_{j=1}^{200}$ denote a set of 200 independent test statistics, such that $T_j \sim \mathcal{N}(\theta_j, 1)$. Given a significance level α, we can construct a confidence interval (CI) for θ_j by the following two-step procedure: (1) evaluate test significance by comparing $|T_j|$ with $z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ stands for the $1-\alpha/2$ quantile of the standard normal distribution (e.g., $z_{1-.05/2} = 1.96$); (2) if $|T_j| \geq z_{1-\alpha/2}$, set the CI as $T_j \pm z_{1-\alpha/2}$; otherwise, regard θ_j as zero without providing any interval. This two-step procedure imitates the common practice of reporting CIs for only those parameters chosen by a model selection algorithm. Benjamini and Yekutieli demonstrated a possible bad

CONTACT Po-Hsien Huang psyphh@mail.ncku.edu.tw Department of Psychology, National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

consequence of this via a simulation with $\alpha = .05$ and a Monte Carlo error less than 0.01. With parameter values $\theta_j = 0$, 0.5, 1, 2, and 4, the corresponding conditional coverage rates were 0, 0.60, 0.84, 0.95, and 0.97. The rates for $\theta_j = 0$, 0.5, and 1 were smaller than their nominal level $1-\alpha = 0.95$. This simulation indicates that this two-step procedure tends to construct too narrow intervals when the parameter value θ_j is small. Benjamini and Yekutieli further showed that the use of Bonferroni adjustment cannot solve this problem. Their adjusted procedure selected θ_j only if $|T_j| \geq z_{1-\alpha/(2\cdot 200)}$, and constructed a Bonferroni corrected interval by $T_j \pm z_{1-\alpha/(2\cdot 200)}$. The corresponding coverage rates became 0, 0.82, 0.97, 1, and 1, still smaller than the nominal level .95 when $\theta_j = 0$ and 0.5.

In the previous example, $T_j \pm z_{1-\alpha/2}$ was constructed only for $|T_j| \geq z_{1-\alpha/2}$, i.e., the statistical inference got conditioned on a selection event. The selection event $|T_j| \geq z_{1-\alpha/2}$ defines a restricted sample space $\{\mathcal{Y} : |T_j| \geq z_{1-\alpha/2}\}$, where $\mathcal{Y} = \{y_n\}_{n=1}^{N}$ is a random sample. Despite that the interval $T_j \pm z_{1-\alpha/2}$ is $1-\alpha$ level for the original or unrestricted sample space $\{\mathcal{Y}\}$, the same might not be true for the restricted sample space. Note that the event $|T_j| \geq z_{1-\alpha/2}$ is not uninformative. When T_j indicates the significance level for testing a parameter θ_j, the event $|T_j| \geq z_{1-\alpha/2}$ corresponds to a sample space that yields a larger estimate for $|\theta_j|$ than the unrestricted sample space. The naive construction of the above CI ignores both the randomness introduced by the selection process and the sample space restriction implied by the chosen model. Consequently, if the true parameter value is close to zero, the naive intervals for the selected parameters are generally too narrow, and tend to indicate significant results. The problem is not only statistical but also substantive. The author believes that ignoring the influence of model selection on statistical inference is one of the potential sources for the reproducibility crisis in psychological science (e.g., Open Science Collaboration, 2015). Note that the term "model selection" used here can stand for any formal or even informal procedure for choosing an "optimal" model, including the *p*-hacking techniques (Simmons, Nelson, & Simonsohn, 2011).

The present article aims to review several state-of-the-art valid postselection inference methods from the statistical literature, and apply these methods to make valid postselection inferences in SEM settings. Some psychometric works recognized the negative impact of model selection uncertainty (e.g., Lubke & Campbell, 2016; Lubke et al., 2017; Preacher & Merkle, 2012), but they didn't discuss how to make valid postselection inferences for individual parameters. We found only Jin and Ankargren (2018) considering this issue formally. Their study approaches the postselection inference problem by the so-called frequentist model averaging (FMA) technique (Hjort & Claeskens, 2003). This technique averages parameter estimates from multiple candidate models to obtain correct inference results.

Although, FMA can yield CIs with desired coverage rates, our experiences reveal that most psychologists make statistical inferences for parameters conditioned on a single selected model. Such type of inference is sometimes called conditional inference (e.g., Lee, Sun, Sun, & Taylor, 2016; Leeb & Pötscher, 2006). The current work focuses on valid methods under the conditional inference framework.

The construction of valid postselection inference methods is influenced by the choice of model selection algorithms. Different algorithms result in their own selection events. Under SEM settings, exploratory model selection algorithms include nested model comparison by sequential likelihood ratio tests (e.g., Steiger, Shapiro, & Browne, 1985), specification search (e.g., Chou & Bentler, 1990; MacCallum, 1986), and the recently developed penalized likelihood (PL) approach (Huang, Chen, & Weng, 2017; Jacobucci, Grimm, & McArdle, 2016).

In this article, we consider two algorithm-independent methods to accomplish the same goal: the data splitting (DS) method (Cox, 1975) and the postselection inference (PoSI) method (Berk, Brown, Buja, Zhang, & Zhao, 2013). In principle, these two methods work with all the mentioned selection algorithms for SEM. However, they tend to be conservative because of their generality. They sometimes result in too wide CIs.

According to recent advances in postselection inference—under LASSO (least absolute shrinkage and selection operator; Tibshirani, 1996) or under L_1-penalized regression—an efficient polyhedral (PH) method can be established to construct CIs (Lee et al., 2016; Taylor & Robert, 2018). Remarkably, when the sampling distribution of the L_1-penalized estimator is derived on the basis of the polyhedral lemma, exact postselection inference can be obtained. Despite this exactness, the PH method is algorithm-specific. It was originally designed for the L_1-penalized model. The author thinks that it might be very challenging to extend the PH method to other selection algorithms for SEM. Therefore, the present study only considers the inference issue after conducting an L_1-penalized

SEM, so that we can compare DS and PoSI with the efficient PH method.

Like other classical subset methods (e.g., forward selection), L_1-penalization automatically generates a set of candidate models with different degrees of complexity—via sparse estimation (see Tibshirani, 1996). After choosing an appropriate penalty level, an optimal model can be determined. The solution path of L_1-penalized estimates across different penalty levels is continuous—making it different from other classical methods. Thus we can characterize the selection event via a simple optimality condition. The PH method is established by this good property.

This article is organized as follows. The next section describes a statistical framework for postselection inference. L_1-penalized SEM and a motivating example section introduces the L_1-penalized SEM and a motivating example. In Valid postselection inference methods section, three valid postselection inference methods, including DS, PoSI, and PH, are introduced. Simulation study section presents a simulation study to evaluate and compare the performance of the reviewed inference methods. A real world data illustration is provided. Finally, merits and limitations of the current work are discussed.

Statistical framework for postselection inference

Let y denote a P-dimensional random vector with a mean vector μ and a covariance matrix Σ. The population moment vector is represented by $\tau = (\mu, \sigma)$, where $\sigma = \text{vech}(\Sigma)$ is a $P(P+1)/2$-dimensional covariance vector and $\text{vech}(\cdot)$ is an operator that stacks the non-duplicated elements of a symmetric matrix. We use $\tau(\theta) = (\mu(\theta), \sigma(\theta))$ to denote an SEM model for the population moment vector $\tau = (\mu, \sigma)$, where θ is a Q-dimensional parameter vector with θ_q being its qth component.

During model selection, both the sparsity pattern and the numerical value of θ should be estimated using a sample data set. By the sparsity pattern of θ, we mean a representation that indicates which elements are nonzero in θ. More specifically, let $\{\mathcal{M}_j\}_{j=1}^{J}$ denote a set of candidate models. We assign each model an index set indicating the nonzero elements of θ. Hence $\mathcal{M}_j$ must be a subset of $\{1, 2, ..., Q\}$. For example, $\mathcal{M}_j = \{1, 2, 3\}$ means that for the jth model $\theta_q = 0$ for $q > 3$. The complexity of $\mathcal{M}_j$ is interpreted through its cardinality $|\mathcal{M}_j|$, counting the elements within. For the most exploratory case, we have $J = 2^Q$ candidate models, considering all possible sparsity patterns of $\theta_1, \theta_2, \ldots, \theta_Q$.

Through a random sample and a model selection procedure, an optimal model $\hat{\mathcal{M}}$ can be obtained. For example, $\hat{\mathcal{M}} = \mathcal{M}_j$ means that the jth model is regarded optimal according to the given data and the selection algorithm. After obtaining $\hat{\mathcal{M}} = \mathcal{M}$, the corresponding parameter estimate $\hat{\theta}_{\mathcal{M}}$ can also be derived. For example, we may use maximum likelihood (ML) to estimate $\theta_{\mathcal{M}}$, the model parameter under the chosen model.

Note that in typical applications of SEM, it is not necessary to consider all sparsity patterns for θ. Substantive theories can be used to exclude some possibilities. For example, in most cases, residual variances are considered nonzero for psychological test data. Without loss of generality, we assume $\theta = (\psi, \phi)$ for some R-dimensional ψ and S-dimensional ϕ. The subvector ψ is formed by all the fixed nonzero and freely estimated parameters across all candidate models. On the other hand, the sparsity pattern of ϕ is unknown in advance. It should be determined by some model selection procedure. Similarly, any parameter under a specific model $\mathcal{M}$ can be partitioned as $\theta_{\mathcal{M}} = (\psi_{\mathcal{M}}, \phi_{\mathcal{M}})$. Despite that ψ is always included in the specified SEM model, its value still depends on the selected model, and hence the subscript $\mathcal{M}$ for ψ cannot be dropped.

The aim of postselection inference is to make a valid statistical inference for $\theta_{\hat{\mathcal{M}}}$ under $\hat{\mathcal{M}} = \mathcal{M}$. The first difficulty in postselection inference is to determine a target parameter, a population quantity that a CI aims to cover. Let $\mathcal{D}(\theta)$ denote the ML fitting function, which measures the discrepancy between $\tau(\theta)$ and τ. Given a random sample $\mathcal{Y} = \{y_n\}_{n=1}^{N}$, the ML fitting function can be written as

$$\mathcal{D}(\theta) = \text{tr}\left[S\Sigma(\theta)^{-1}\right] - \log|S\Sigma(\theta)^{-1}| - P + \left[m - \mu(\theta)\right]^T \Sigma(\theta)^{-1}\left[m - \mu(\theta)\right], \tag{1}$$

where m is the sample mean vector and S is the sample covariance matrix. Berk et al. (2013) distinguished two perspectives for defining a target parameter in a model selection setting. According to the *full model view*, we quantify the population value of $\theta_{\hat{\mathcal{M}}}$ by

$$\theta^* = \text{argmin}_{\theta \in \Theta} \mathbb{E}[\mathcal{D}(\theta)], \tag{2}$$

where Θ is the parameter space of θ. An advantage of using θ^* is model independence. Regardless of the realization of $\hat{\mathcal{M}}$, the same target is inferred. However, from the quasi-ML theory of White (1982), we expect that θ^* is not the limit point of $\theta_{\hat{\mathcal{M}}}$ for

every realization of $\hat{\mathcal{M}}$. According to the *submodel view*, we define the population target by

$$\theta^*_{\mathcal{M}} = \text{argmin}_{\theta\in\Theta_{\mathcal{M}}}\mathbb{E}[\mathcal{D}(\theta)], \tag{3}$$

where $\Theta_{\mathcal{M}}$ is the parameter space under the selected model $\mathcal{M}$. Clearly, $\theta^*_{\mathcal{M}}$ is just a quasi-true parameter under $\mathcal{M}$. Hence $\hat{\theta}_{\mathcal{M}}$ is statistically consistent to $\theta^*_{\mathcal{M}}$ for every realization of $\hat{\mathcal{M}}$ (see White, 1982). Under the submodel view, the target to be inferred changes with the selected model. Berk et al. (2013) argued that the submodel view corresponds to a more reasonable target than the full model view. In fact, the validity of inference methods described in the next section can be only justified under the submodel view.

To better understand the concept of target parameters, we present a linear regression example with three covariates. Suppose that (1) the three covariates x_1, x_2, and x_3 are standardized with pairwise correlation $\rho = 0.3$; and (2) the response variable y is derived as $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ with $\beta_1 = 0.5$ and $\beta_2 = \beta_3 = 0$, where ϵ is a residual term with variance 0.75. When the chosen model correctly identifies β_1 as nonzero (we call it a correct model), both the full model view and the submodel view quantify the populations of $(\beta_1, \beta_2, \beta_3)$ as $(0.5, 0, 0)$, consistently with the value generating y. However, if only β_2 is regarded nonzero by the chosen model (we call it an incorrect model), the targets yielded by the two perspectives are different. The full model view still considers $(0.5, 0, 0)$ as the population values of $(\beta_1, \beta_2, \beta_3)$, but the submodel view quantifies the population targets as $(0, 0.15, 0)$ after minimizing in Equation (3). Despite that $(0, 0.15, 0)$ seems to be irrelevant to the generation of y, the predicted value $\hat{y} = 0.15 \times x_2$ is still the best approximation for y if only x_2 is used for prediction. Hence, even if a chosen model is incorrect, it is still meaningful to make an inference for the corresponding target quantified by the submodel view.

Our goal is to construct a valid $1-\alpha$ level CI for each nonzero component of $\theta_{\mathcal{M}}$ after obtaining $\hat{\mathcal{M}} = \mathcal{M}$. In particular, for each realization of $\hat{\mathcal{M}}$ and $q \in \hat{\mathcal{M}}$, we hope to construct a random interval $C_{\mathcal{M},q}$ that satisfies

$$\mathbb{P}\left(\theta^*_{\mathcal{M},q} \in C_{\mathcal{M},q} | \hat{\mathcal{M}} = \mathcal{M}\right) \geq 1-\alpha, \tag{4}$$

where $\theta^*_{\mathcal{M},q}$ is the qth element of $\theta^*_{\mathcal{M}}$. Three important things concerning Equation (4) should be noted. (1) The coverage property is conditioned on the event that a particular model $\mathcal{M}$ is selected. In other words, the randomness of $C_{\mathcal{M},q}$ is restricted to the sample space $\{\mathcal{Y}|\hat{\mathcal{M}} = \mathcal{M}\}$. Note that $\hat{\mathcal{M}}$ is a function of the random data $\mathcal{Y}$. (2) Only elements in $\mathcal{M}$ are possible to be inferred. Hence, we cannot construct CI for a parameter that is not chosen by the model selection procedure. From Berk et al. (2013)'s perspective, a non-chosen parameter is thought about as nonexistent in the selected model. (3) Equation (4) states that the conditional coverage rate of $C_{\mathcal{M},q}$ with respect to $\theta^*_{\mathcal{M},q}$ is at least $1-\alpha$. This statement is similar to the definition of the usual $1-\alpha$ level CI (e.g., Casella & Berger, 2002), except that the coverage property is conditioned on the selection event.

L_1-penalized SEM and a motivating example

The current study only considers selection events made by L_1-penalization. In this section, the L_1-penalized SEM is briefly described, see Huang et al. (2017) for further details. Additionally, we present a motivating example to show how the L_1-penalized SEM is conducted. This example will be also used in the next section to demonstrate the use of valid postselection inference methods.

The L_1-penalized SEM is a method that simultaneously selects non-null parameters, and estimates their values in an SEM model. The method is based upon the following penalized likelihood (PL) objective function

$$\mathcal{U}(\theta, \lambda) = \mathcal{D}(\theta) + \lambda||\phi||_1, \tag{5}$$

where λ is a regularization parameter that controls the penalty level, and $||\cdot||_1$ denotes the L_1 norm of a vector, i.e., $||\phi||_1 = \sum_{s=1}^{S}|\phi_s|$. For a fixed λ, a PL estimate, denoted by $\hat{\theta} \equiv \hat{\theta}_{\mathcal{M}_\lambda}$, is defined as a minimizer of $\mathcal{U}(\theta, \lambda)$. Because L_1 penalty can result in a sparse estimate (i.e., an estimate with some zero elements), the corresponding model is presented as $\mathcal{M}_\lambda = \{q|\hat{\theta}_q(\lambda) \neq 0\}$. Thereby $\mathcal{M}_\lambda$ is the set containing the indices of nonzero elements in $\hat{\theta}(\lambda)$.

Let $T_\lambda = N \times \mathcal{D}(\hat{\theta})$ denote the test statistic under λ. Given a set of penalty levels $\Lambda = \{\lambda_j\}_{j=1}^{J}$, an optimal penalty level can be selected by finding a $\hat{\lambda} \in \Lambda$ that minimizes either the Akaike information criterion (AIC; Akaike, 1974)

$$AIC_\lambda = T_\lambda - 2df_\lambda, \tag{6}$$

or the Bayesian information criterion (BIC; Schwarz, 1978)

$$BIC_\lambda = T_\lambda - \log(N)df_\lambda, \tag{7}$$

where df_λ denotes the degrees of freedom for the model determined by λ. In most cases, the degrees of freedom can be calculated by $df_\lambda = P(P+3)/2 - e_\lambda$ with e_λ being the number of nonzero and estimated elements in $\hat{\theta}$. It is well known that BIC results in

consistent model selection while AIC has an overfitting tendency (e.g., Bozdogan, 1987; Huang, 2017; Huang et al., 2017). We may also consider other types of information criteria (see Bollen, Harden, Ray, & Zavisca, 2014; Lin, Huang, & Weng, 2017, for reviews). For example, the Haughton's BIC (HBIC; Haughton, 1988) is defined as

$$HBIC_\lambda = T_\lambda - \log\left(\frac{N}{2\pi}\right) df_\lambda. \quad (8)$$

Under classical model selection settings, some simulations showed the advantage of HBIC over AIC and BIC in terms of choosing the true model (e.g., Bollen et al., 2014; Haughton, Oud, & Jansen, 1997; Lin et al., 2017).

After choosing $\hat{\lambda}$, the corresponding optimal model is $\mathcal{M}_{\hat{\lambda}} = \{q|\hat{\theta}_q(\hat{\lambda}) \neq 0\}$, and the final PL estimate is written as $\hat{\theta}_{\hat{\mathcal{M}}} \equiv \hat{\theta}_{\mathcal{M}_{\hat{\lambda}}}$. Because the L_1-penalized estimator $\hat{\theta}_{\hat{\mathcal{M}}}$ is biased, we further calculate a debiased version for $\hat{\theta}_{\hat{\mathcal{M}}}$, and use the debiased estimator $\tilde{\theta}_{\hat{\mathcal{M}}}$ to construct CIs. The simplest way to obtain $\tilde{\theta}_{\hat{\mathcal{M}}}$ is to calculate the unpenalized ML estimate under $\hat{\mathcal{M}} = \mathcal{M}_{\hat{\lambda}}$. Another way is to consider a one-step estimate (e.g., van de Geer, Bühlmann, Ritov, & Dezeure, 2014; Zhang & Zhang, 2014) (see the Appendix for the derivation). Theoretically, the one-step estimator has the same asymptotic distribution as the corresponding unpenalized ML estimator.

After obtaining a selected model, the naive method uses the trivial CIs for each selected parameter. This method assumes that the chosen model is predetermined. Given $\hat{\mathcal{M}} = \mathcal{M}$, the naive method constructs the interval $C^N_{\mathcal{M},q}$ as

$$C^N_{\mathcal{M},q} = \left[\tilde{\theta}_{\mathcal{M},q} - z_{1-\alpha/2} \times s.\hat{e}.\left(\tilde{\theta}_{\mathcal{M},q}\right), \tilde{\theta}_{\mathcal{M},q} + z_{1-\alpha/2} \times s.\hat{e}.\left(\tilde{\theta}_{\mathcal{M},q}\right)\right], \quad (9)$$

where z_α is the α-quantile of the standard normal distribution, and $s.\hat{e}.(\tilde{\theta}_{\mathcal{M},q})$ is an estimated standard error for $\tilde{\theta}_{\mathcal{M},q}$. The standard error $s.\hat{e}.(\tilde{\theta}_{\mathcal{M},q})$ can be obtained by either inverting the Fisher information matrix, or by using a sandwich formula (see Yuan & Hayashi, 2006). According to the author's experience, only the naive method is used in SEM practice. However, the naive method is generally incorrect because it ignores model selection in the process of data analysis. The naive CI might be too narrow, and thus it tends to yield an empirical coverage rate smaller than $1-\alpha$.

Now, we consider an example of L_1-regularized factor analysis (e.g., Hirose & Yamamoto, 2015; Huang et al., 2017). The data set—collected by Holzinger and Swineford (1939)—contains the responses of 301 seventh- and eighth-grade students responding to 24 psychological tests, of which only the first 19 were used for our analysis. These tests included *visual perception* (y_1), *cubes* (y_2), *paper form board* (y_3), *flags* (y_4), *general information* (y_5), *paragraph comprehension* (y_6), *sentence completion* (y_7), *word classification* (y_8), *word meaning* (y_9), *addition* (y_{10}), *code* (y_{11}), *counting groups of dots* (y_{12}), *straight and curved capitals* (y_{13}), *word recognition* (y_{14}), *number recognition* (y_{15}), *figure recognition* (y_{16}), *object number* (y_{17}), *number-figure* (y_{18}), and *figure-word* (y_{19}). These 19 tests were assumed to be indicators of four oblique factors: spatial (f_1), verbal (f_2), speed (f_3), and memory (f_4). The major indicators for f_1, f_2, f_3, and f_4 were $y_1-y_4, y_5-y_9, y_{10}-y_{13}$, and $y_{14}-y_{19}$, respectively. The loadings of the major indicators were specified as freely estimated parameters. Other loadings were still estimated, but penalized by an L_1 regularizer. Because the variances of the tests were quite different, they needed to be standardized for further analysis. An optimal penalty level was chosen from a candidate set $\Lambda = \{\lambda_j\}_{j=1}^{100}$ ranging from 0.01 to 1.0. The construction of Λ was the same as that in our simulation (see the section of simulation study).

Based on the value of AIC, an optimal penalty level $\hat{\lambda} = 0.089$ was obtained. Under this penalty level, 29 of the 57 penalized loadings were identified as nonzero. The central task of postselection inference is to make statistical conclusions for these selected loadings, as well as other freely estimated parameters. Figure 1 presents the confidence intervals for the selected loadings. These intervals were constructed by the commonly used naive method and the three postselection inference methods that will be introduced in the next section. For the zero loading estimates, no intervals were provided. The naive intervals indicated that 12 selected loadings by L_1 differed from zero significantly. However, the naive method was usually too liberal, as indicated by our simulation (see the section of simulation study), and hence inferences based on naive CIs could not be trusted.

Valid postselection inference methods

In this section, three valid postselection CI methods are introduced: the data splitting, the postselection inference, and the polyhedral method.

Data splitting

During data splitting (DS), a sample data set is split into two disjoint parts. The first part is used to choose

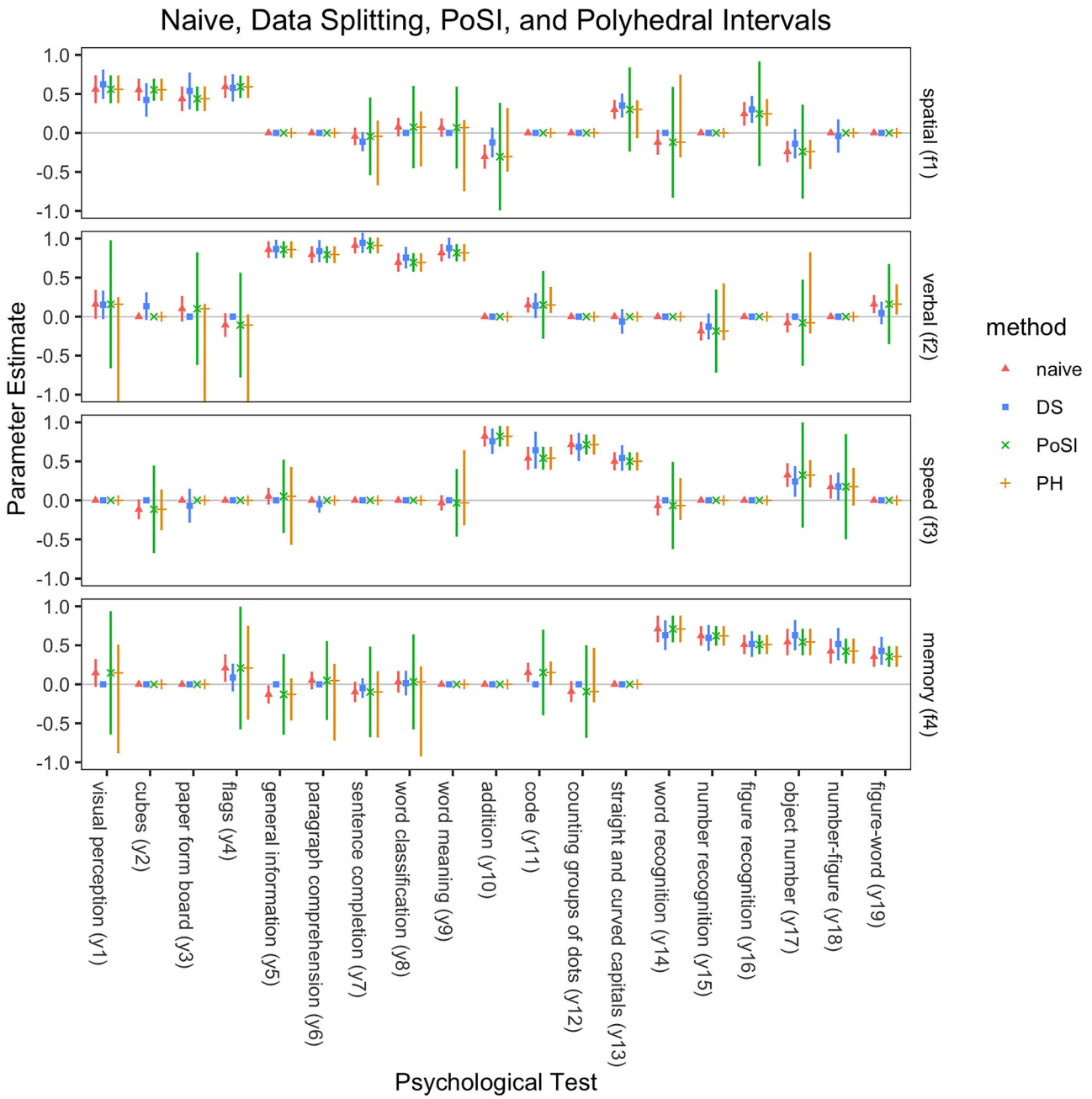

Figure 1. Naive, data splitting (DS), postselection inference (PoSI), and polyhedral (PH) confidence intervals ($\alpha = 0.05$) for the first 19 psychological tests of Holzinger & Swineford (1939). The point estimates are represented by the points within the intervals.

an optimal model, and the second part is used to construct CIs conditioned on the chosen model (Cox, 1975). Technically, we split the sample data $\mathcal{Y}$ into $\mathcal{Y}^A$ and $\mathcal{Y}^B$, such that $\mathcal{Y}^A \cup \mathcal{Y}^B = \mathcal{Y}$ and $\mathcal{Y}^A \cap \mathcal{Y}^B = \emptyset$. An optimal model $\hat{\mathcal{M}} = \mathcal{M}$ is obtained from the starting point $\mathcal{Y}^A$. Then, $C^{DS}_{\mathcal{M},q}$ is constructed by Equation (9) using $\tilde{\theta}_{\mathcal{M},q}$ and $\hat{s.e.}(\tilde{\theta}_{\mathcal{M},q})$ estimated by an unpenalized ML through $\mathcal{Y}^B$. Even though, DS looks very simple, it results in valid CIs regardless of the implemented selection procedure. The main drawback of DS is obvious: a smaller sample size of the partitioned data makes the selection process unstable and the CIs wider.

To apply the DS in our motivating example, we splitted the original data set into two parts. Through the first part of data, an optimal model $\mathcal{M}_{\hat{\lambda}}$ was chosen. Under $\mathcal{M}_{\hat{\lambda}}$, 19 of the 57 penalized loading estimates were identified as nonzero. Then we fitted the chosen model to the second part of data with unpenalized ML. The DS intervals for the chosen loadings were simply the naive CI calculated by using the second part of data. The result showed that only

$f_1 \rightarrow y_{13}, f_1 \rightarrow y_{16}$, and $f_3 \rightarrow y_{17}$ were statistically significant (see Figure 1).

Postselection inference

Postselection inference (PoSI) tries to find a multiplier that can be legitimately used for each selected $\mathcal{M}$ (Berk et al., 2013). Formally, the method searches for a constant k_α, such that

$$\mathbb{P}\left(\max_{\mathcal{M}} \max_{q\in\mathcal{M}} \left|\frac{\tilde{\theta}_{\mathcal{M},q}-\theta^*_{\mathcal{M},q}}{s.\hat{e}.\left(\tilde{\theta}_{\mathcal{M},q}\right)}\right| \le k_\alpha\right) \ge 1-\alpha. \tag{10}$$

After deriving k_α, the PoSI interval $C^{PoSI}_{\mathcal{M},q}$ is constructed via

$$C^{PoSI}_{\mathcal{M},q} = \left[\tilde{\theta}_{\mathcal{M},q} - k_\alpha \times s.\hat{e}.\left(\tilde{\theta}_{\mathcal{M},q}\right), \tilde{\theta}_{\mathcal{M},q} + k_\alpha \times s.\hat{e}.\left(\tilde{\theta}_{\mathcal{M},q}\right)\right]. \tag{11}$$

The definition of k_α in Equation (10) implies that $C^{PoSI}_{\mathcal{M},q}$ satisfies the conditional coverage property in Equation (4). An evident PoSI constant can be obtained by Scheffe's method: $k_\alpha = \sqrt{\chi^2_{S,1-\alpha}}$, where $\chi^2_{S,1-\alpha}$ is the $1-\alpha$-quantile of the χ^2 distribution with S degrees of freedom.[1] Scheffe's PoSI constant is generally too large, and it makes for a conservative CI. For linear regression problems, Berk et al. (2013) suggested a numerical method to compute a better PoSI constant, but this method is time consuming. For instance, several examples in the R package PoSI (Buja & Zhang, 2017) require 10+ minutes to finish. In addition, the development of the numerical method in Berk et al. (2013) depends on the structure of the linear regression problem. It might be difficult to extend the numerical method to SEM cases. Therefore, the present study simply uses Scheffe's PoSI constants to construct CIs for SEM model parameters.

To calculate the PoSI intervals for the selected loadings, we first derived the Scheffe's PoSI constant by $k_\alpha = \sqrt{\chi^2_{57,.95}} = 8.70$. Then the PoSI intervals were obtained by using Equation (11) with $k_\alpha = 8.70$. Figure 1 indicated that none of the selected loadings were recognized as significant by PoSI. As we shall see in the section of simulation study, PoSI generally yields conservative inference results. When using R package lslx to conduct L_1-penalized SEM (Huang, in press), PoSI intervals can be easily obtained by setting inference="scheffe" in the summarize() method.

[1]In linear regression problems with S covariates to be chosen, the Scheffe's PoSI constant is $k_\alpha = \sqrt{S \times F_{S,N-S,1-\alpha}}$, where $F_{S,N-S,1-\alpha}$ is the $1-\alpha$-quantile of the F distribution with degrees of freedom S and $N - S$. In SEM settings, we should consider $S \times F_{S,N-S,1-\alpha}$ with $N \rightarrow \infty$, which converges to $\chi^2_{S,1-\alpha}$.

The polyhedral method

The polyhedral (PH) method derives the sampling distribution of the one-step debiased estimate under a selection event made by L_1-penalization with a fixed penalty level (Lee et al., 2016; Taylor & Robert, 2018). Because the sampling distribution is known, the CIs for the selected parameters can be constructed. Details of the PH method can be found in the Appendix. To describe the sampling distribution, let $\vartheta_{\mathcal{M}}$ denote a subvector of $\theta_{\mathcal{M}}$ formed by $\{\theta_{\mathcal{M},q}\}_{q\in\mathcal{M}}$, i.e., $\vartheta_{\mathcal{M}}$ only includes the nonzero elements of $\theta_{\mathcal{M}}$. Furthermore, we use $\varphi_{\mathcal{M}}$ to denote a subvector formed by the selected elements $\phi_{\mathcal{M}}$, which implies that $\vartheta_{\mathcal{M}} = (\psi_{\mathcal{M}}, \varphi_{\mathcal{M}})$. Under a fixed penalty level, the large sample distribution of the one-step estimator $\tilde{\vartheta}_{\mathcal{M}}$ equals

$$\sqrt{N}\left(\tilde{\vartheta}_{\mathcal{M}} - \vartheta^*_{\mathcal{M}}\right) \sim \mathcal{N}\left(0, \hat{\mathcal{C}}_{\mathcal{M}}\right), \tag{12}$$

being restricted to the event

$$\left\{\text{sign}\left(\tilde{\varphi}_{\mathcal{M}} - \mathcal{P}_{\mathcal{M}}\hat{\mathcal{F}}^{-1}_{\mathcal{M}}\begin{pmatrix}0\\ \lambda\hat{s}_{\mathcal{M}}\end{pmatrix}\right)\right\} = \hat{s}_{\mathcal{M}}, \tag{13}$$

where $\hat{\mathcal{C}}_{\mathcal{M}}$ is an estimated covariance matrix of $\tilde{\vartheta}_{\mathcal{M}}$, $\hat{\mathcal{F}}_{\mathcal{M}}$ is the observed Fisher information matrix with respect to $\mathcal{M}$, $\hat{s}_{\mathcal{M}}$ is the sign vector of $\tilde{\varphi}_{\mathcal{M}}$, i.e., $\hat{s}_{\mathcal{M}} = \text{sign}(\tilde{\varphi}_{\mathcal{M}})$, and $\mathcal{P}_{\mathcal{M}}$ is a projection matrix that selects rows corresponding to $\tilde{\varphi}_{\mathcal{M}}$. According to the polyhedral lemma, each element of $\tilde{\vartheta}_{\mathcal{M}}$ is asymptotically distributed as a truncated normal variate. A PH interval $C^{PH}_{\mathcal{M},q}$ can be obtained by inverting the cumulative distribution function of the corresponding truncated normal variate. Under linear regression and a fixed penalty level, $C^{PH}_{\mathcal{M},q}$ is exact, which means that the coverage rate of a PH interval is exactly $1-\alpha$. Under a general L_1-regularized estimation, $C^{PH}_{\mathcal{M},q}$ is only asymptotically exact. However, the validity of PH intervals is only ensured for fixed penalties. The value of the regularization parameter must be predetermined, and it cannot be tuned by any data-driven method. Nevertheless, existing simulation shows that $C^{PH}_{\mathcal{M},q}$ still performs well when an optimal penalty level is chosen post hoc (Taylor & Robert, 2018). Another drawback of PH is that the truncated normal probability is hard to evaluate. Occasionally, the PH method results in an infinite interval. In our simulation, about 0.2% of the PH intervals were infinite.

It might not be a trivial task to write a code for calculating PH CIs. Fortunately, they can be also obtained by using lslx package (with inference="polyhedral" in the summarize() method). For the motivating example, the PH CIs showed that only five selected loadings were thought to be nonzero

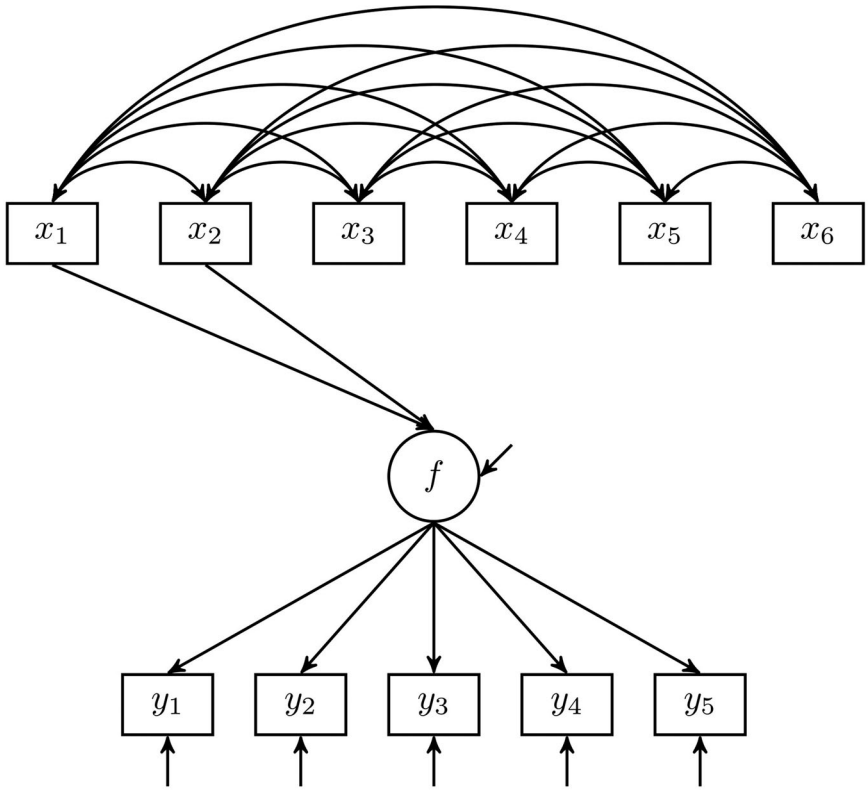

Figure 2. The small MIMIC model for generating data. It includes five indicators (y_1 – y_5), one latent factors (f), and six potential causes (x_1 – x_6). The non-null path coefficient is set to 0.1, 0.2, or 0.3, depending on the simulation condition. All factor loadings and correlations among causes are set to 0.7 and 0.2, respectively. The values of residual variances are so chosen as to let the observed variables and latent factors have unit variances.

statistically, including $f_1 \rightarrow y_{16}, f_1 \rightarrow y_{17}, f_2 \rightarrow y_{11}, f_2 \rightarrow y_{19}$, and $f_3 \rightarrow y_{17}$ (see Figure 1).

Simulation study

In this section, we compare the empirical performances of $C^{N}_{\mathcal{M},q}, C^{DS}_{\mathcal{M},q}, C^{PoSI}_{\mathcal{M},q}$, and $C^{PH}_{\mathcal{M},q}$. The simulation evaluates the postselection CIs for path coefficients with the help of the multiple indicators and multiple causes (MIMIC) model (Jöreskog & Goldberger, 1975). Four factors are considered here: the size of the models (small and large), the size of the non-null effects (small, medium, and large), the sample sizes (100, 200, 400, 600, and 800), and the type of selectors (AIC, BIC, and HBIC).

The size of the models reflects different analytical settings. Figure 2 shows the small size MIMIC model, which includes six potential causes, five indicators, and one latent factor. Only the first two causes had non-null effects on the latent factor. For the large size model, the true model specification was similar to that of the small one, except that (1) 12 potential causes were considered; and (2) the first four causes were non-null. Because the success of using information criteria for choosing relevant features depends on the magnitude of non-null effects (e.g., Lin et al., 2017; Vrieze, 2012), the values of non-null effects were set as 0.1, 0.2, or 0.3 for representing the small, medium, or large effect.[2] For each combination of the model sizes and the effect sizes, we generated data for sample sizes of 100, 200, 400, 600, and 800 from multivariate normal distribution with the specified covariance matrix.

Each data set was analyzed under its corresponding MIMIC model having all path coefficients estimated with L_1-penalization. For each analysis, the corresponding optimal penalty level was selected from $\Lambda = \{\lambda_j\}_{j=1}^{60}$ by either AIC, BIC, or HBIC. Here, $\lambda_j = \exp\{b_j\}$ with b_j being the jth element of B, a set of equally spaced values between $\log(0.01)$ and $\log(0.60)$.

[2]Despite that we used" large" to qualify the value of 0.3, a standardized regression coefficient $\beta = 0.3$ only represents a medium effect according to Cohen (1992). Thus, the values of the non-null effects are considered only small to medium. We just used the terms "small", "medium", and "large" for convenience. When we adopted a large effect defined by Cohen here (e.g., $\beta = 0.5$), the resulting R^2 became larger than one, which is impossible.

By the construction of Λ, the candidates for the tuning parameters were in log-scale (see Bühlmann & van de Geer, 2011, p. 38). Our experiment indicated that (1) $\lambda_{60} = 0.60$ was large enough to shrink all regression coefficients to zero; (2) the grid formed by the elements of Λ was dense enough to approximate the solution paths of the PL estimates. After obtaining a selected model, $C^{N}_{\mathcal{M},q}, C^{DS}_{\mathcal{M},q}$, $C^{PoSI}_{\mathcal{M},q}$, and $C^{PH}_{\mathcal{M},q}$ were constructed for each nonzero path coefficient. The significance level α was set to 0.05. Except for $C^{DS}_{\mathcal{M},q}$, all CIs were calculated based on the one-step estimator.

In summary, there were a total of $2 \times 3 \times 5 \times 3 = 90$ conditions. For each condition, 2000 successful replications were submitted for analysis. A replication was considered successful if the PL estimate was derived from a convergent optimization.[3] The simulations were conducted within the R environment (R Core Team, 2018). Package lslx (Huang, in press) was used to implement the L_1-penalized SEM and the considered inference methods.

Two indices were used to evaluate the performance of the considered intervals: coverage rate (CR) and interval length (IL). For any single parameter $\theta_{\mathcal{M},q}$, the CR of $C_{\mathcal{M},q}$ was defined as the proportion of the time that the interval contained the corresponding population targets across all replications, given that $\theta_{\mathcal{M},q}$ was selected. Here, the population target was the quasi-true parameter under the chosen model (see Equation (3)). In an ideal case, a valid method would result in CIs with CRs equal to $1-\alpha$. If a method only yields CRs larger than $1-\alpha$, it is still valid according to the definition by Equation (4), and we would say that the method constructed CIs with controlled CRs. The IL of $C_{\mathcal{M},q}$ is the median of the length of all intervals for $\theta_{\mathcal{M},q}$. Among several valid methods, a method is considered the most efficient if it results in a CI with the narrowest IL. Because it is difficult to present CR and IL for each model-dependent parameter,[4] we only presented the average CR and the median IL for two categories of parameters: nonzero targets and zero targets. The nonzero targets are path coefficients associated with non-null causes, while the zero targets are coefficients for null causes under the true model. In typical cases, the nonzero targets are parameters that researchers are interested in.

According to the theory of these postselection inference methods, we expected that (1) the PH method would perform the best with a controlled CR and a relatively narrow IL. In addition, its performance would improve with sample size; (2) both DS and PoSI could control CR, but would tend to be conservative with wide IL; (3) the naive method would generally yield a too small CR by ignoring the presence of model selection, even if its IL was the narrowest.

Before presenting the CR and the IL results, we examine the positive rates (PR) for selecting the nonzero targets and the zero targets under each condition (see Figure 3). The PR result shows that nonzero targets can be chosen consistently under large effect sizes. For small and medium effects, PR increases with sample size. However, BIC and HBIC could not properly select nonzero targets of value 0.1, even with a sample size of 800. DS uses half of the sample data for model selection. Hence, the half-data selection result is not as good as the full-data result in terms of the PR for nonzero targets. In general, our PR result is consistent with the typical behavior of information criteria found in SEM literature (e.g., Lin et al., 2017; Vrieze, 2012).

Figures 4 and 5 show the CR and the IL results of the simulation. In most conditions, the PH method performed well with controlled CR and relatively narrow IL, but it could not control CR for zero targets under BIC and HBIC combined with small effects. The fact that the CR of PH intervals was not always close to the nominal level $1-\alpha = 0.95$ is probably due to the randomness introduced by penalty level selection. It is worth noting that the IL of PH intervals for nonzero targets was quite similar to the IL of those in the naive method under medium and large effects. This means that the efficiency loss of PH is negligible for large nonzero targets. DS performed surprisingly well. It yielded CR values very close to 0.95. The IL of DS intervals was not the narrowest compared with other valid methods, but DS performed reasonably well and stable for both types of targets. PoSI could also yield $1-\alpha$ level CIs. However, the CR of PoSI intervals was close to 1 across almost all conditions, which shows that PoSI is too conservative. PoSI produced the widest intervals. CR for the naive method was only controlled for nonzero targets under some AIC conditions. In general, the naive method yielded too small CR, despite that its IL was the smallest. The naive method performed worse for zero targets under BIC/HBIC than under AIC.

[3]The five-number summary for the rates of convergent solutions was 27%, 93%, 100%, 100%, and 100%. We found that nonconvergent solutions mainly occurred when both the model size and the number of non-null effects were large and the sample size was 100. For other conditions, the rate of convergent solutions exceeded 93%.

[4]Recall that the population target is model-dependent. Suppose that there are Q potential causes. Without the duplicated cases, there are $Q \cdot 2^{Q-1}$ model-dependent parameters since: (1) Q path coefficients are considered; and (2) each potential cause appears in 2^{Q-1} candidate models.

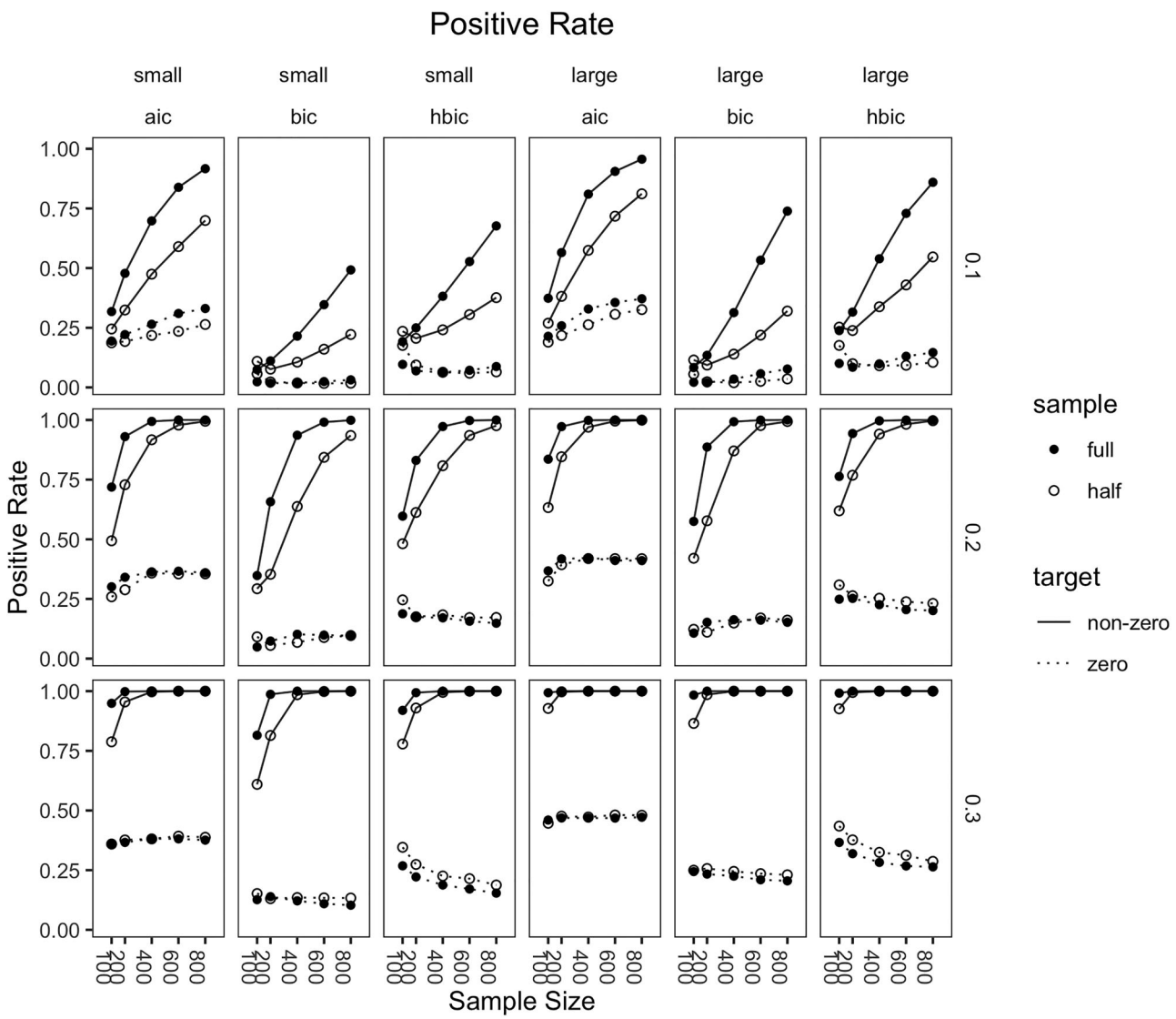

Figure 3. Positive rates of full-data selection and half-data selection for nonzero and zero target parameters under AIC, BIC, and HBIC selectors across effect sizes, model sizes, and sample sizes.

Why? It is well known that BIC/HBIC favors a parsimonious model because of its heavier penalty for complexity (i.e., the term $\log(N)$ for BIC in Equation (7)). Any BIC/HBIC selected parameter reduces the discrepancy function substantially, which implies that the parameter appears" statistically significant". As a result, the constructed CI for a zero target also shifted from zero toward its parameter estimate.

Example: psychological domain of WHOQOL data

The example applies the inference methods to a MIMIC model for exploring measurement invariance. We adopted the data set from Chen and Yao (2015). It contains the responses of 158 males and 240 females to six items from the Taiwan Version of WHOQOL-BREF (World Health Organization Quality of Life—Short Form; The WHOQOL Taiwan Group, 2005). The six items are *positive feeling*, *spirit*, *thinking*, *body image*, *self-esteem*, and *negative feeling*. These items were used to measure the psychological domain of QOL, using the five-point Likert type scale. An L_1-penalized MIMIC model was specified to probe the intercept invariance across males and females. The loading of *positive feeling* was fixed at 1 for scale setting. Other loadings were freely estimated. The regression coefficients from gender to each item were set as penalized parameters. Based on the value of AIC, full-data selection chose $\hat{\lambda} = 0.051$ from the same Λ set that we used in the previous example. The final model indicated that only the regression coefficient for the fifth item *self-esteem* was nonzero ($\hat{\beta}_2 = 0.138$). From

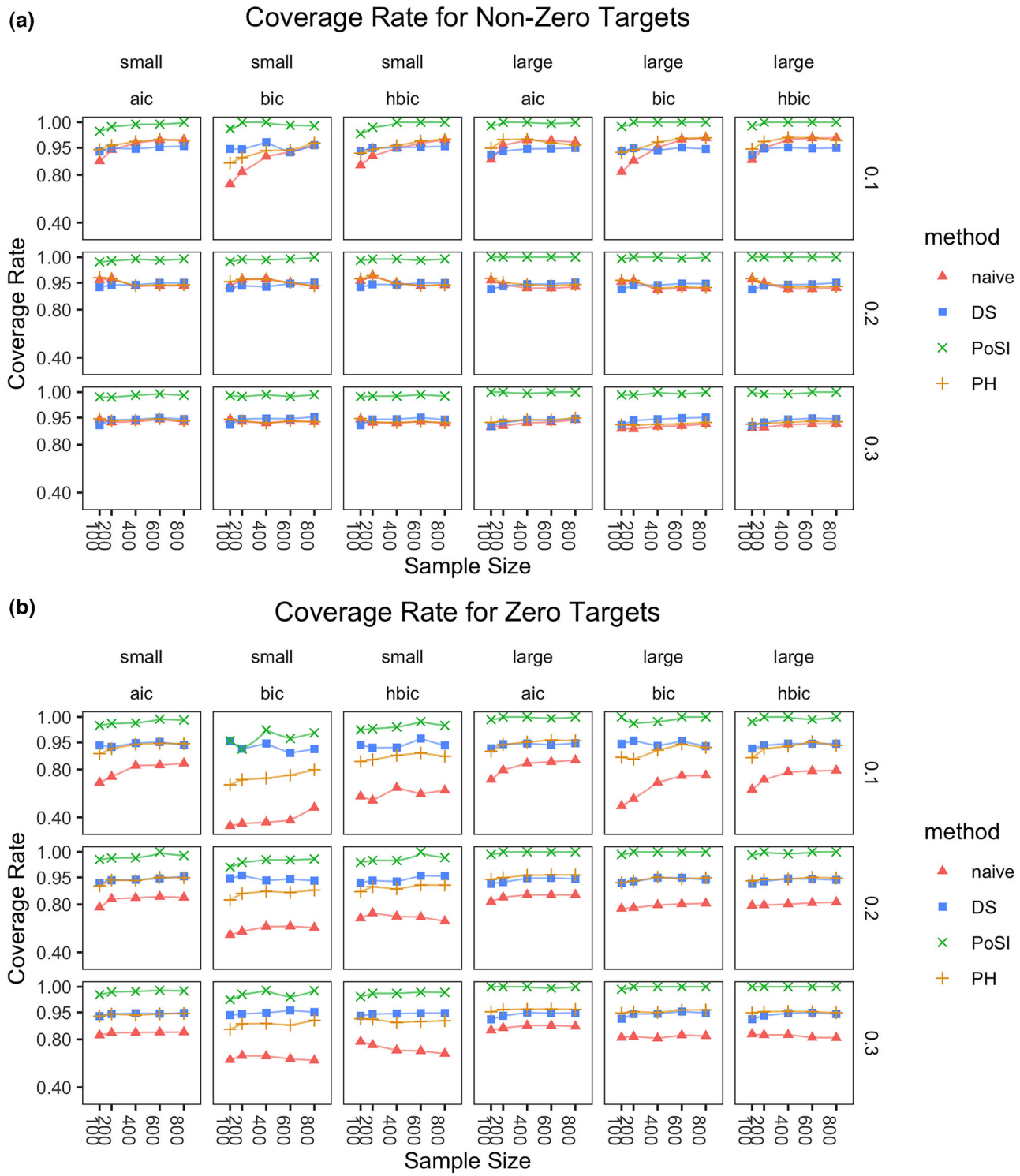

Figure 4. Coverage rates of the naive, data splitting (DS), postselection inference (PoSI), and polyhedral (PH) confidence intervals for nonzero and zero target parameters under AIC, BIC, and HBIC selectors across effect sizes, model sizes, and sample sizes. Note that the *y*-axis is scaled by $tanh^{-1}$.

the naive CI ($C^N = [0.009, 0.266]$), the intercept of *self-esteem* appeared non-invariant. However, from both PoSI ($C^{PoSI} = [-0.096, 0.371]$) and the polyhedral interval ($C^{PH} = [-0.053, 0.266]$), all intercepts appeared invariant across males and females. The half-data selection result made by AIC indicated that no penalized coefficients were identified as nonzero ($\hat{\lambda} = 1$). Despite the selection results yielded by full-data and half-data procedures were different, the final conclusion about the intercept invariance were the same.

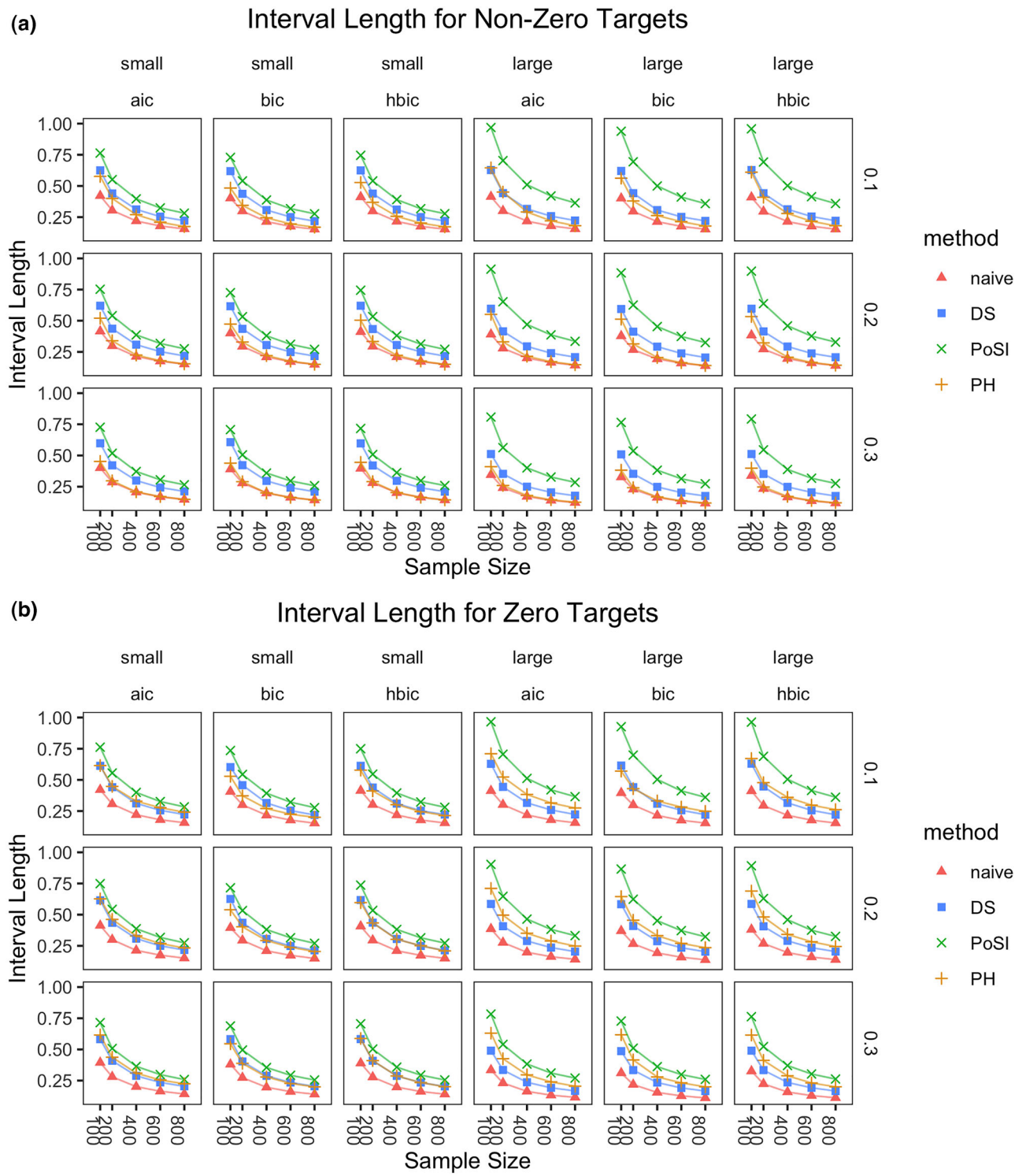

Figure 5. Length of naive, data splitting (DS), postselection inference (PoSI), and polyhedral (PH) confidence intervals for nonzero and zero target parameters under AIC, BIC, and HBIC selectors across effect sizes, model sizes, and sample sizes.

Discussion

The present study reviewed three valid postselection inference methods, and applied them to structural equation modeling (SEM). Therein, data splitting (DS) and postselection inference (PoSI) were algorithm-independent, while the polyhedral (PH) method was originally designed for L_1-penalized methods. A numerical experiment was conducted to compare the performances of the three valid methods with that of the naive approach. As expected, the naive method failed to construct valid intervals, and the three valid methods worked under most simulated conditions

with their own pros and cons. Two real world data examples showed that the valid methods can yield different conclusions from that of the naive method. In general, the valid methods were more conservative to protect the type I error rate.

The practical implications of the present study are as follows. (1) The naive method should never be used after model selection. Although the simulation showed that the naive method sometimes produces good intervals for nonzero targets, in general, its CR for zero targets is always too small. Therefore, the naive method tends to obtain significant results for selected zero targets. It seems to the author that the wide use of the naive method after model selection results in numerous false positive findings in psychology, which is unfortunate. (2) The DS method is generally recommended. It is very simple, and it can yield reasonably good intervals with excellent CR control. DS can also be used for cross-validation (CV) (e.g., Camstra & Boomsma, 1992; Cudeck & Browne, 1983; MacCallum, Roznowski, Mar, & Reith, 1994). Although CV after model selection has been advocated by some researchers (e.g., Browne, 2000; MacCallum, Roznowski, & Necowitz, 1992), its use is only occasionally reported (Jackson, Gillaspy, & Purc-Stephenson, 2009; MacCallum & Austin, 2000). The existing CV methods in SEM are mainly used for overall model evaluation, not for individual parameter testing. The author would suggest the regular practice of using DS for statistical inference after any type of model selection. (3) The PH method should be implemented with care. The simulations indicated that PH performs well when AIC is used. In addition, PH results in efficient intervals for medium and large nonzero parameters, which are the targets that most studies wish to discover. (4) The PoSI method with Scheffe's constant is only recommended if researchers tend to take a conservative stance or the sample size is large.

Although the current article only considers model selection via sparse estimation made by L_1-penalization—except for the polyhedral-related results—the author believes that the findings can be generalized to any data-driven method for model selection. The two algorithm-independent methods can be applied to most model selection procedures in SEM with only slight adjustment. Of course, to enhance our understanding of these postselection inference methods, future studies can directly evaluate their performance under different selection algorithms and simulated settings.

Recently, many L_1-penalized estimation methods have been developed in psychometrics (Chen, Liu, Xu, & Ying, 2015; Hirose & Yamamoto, 2014, 2015; Huang et al., 2017; Jacobucci et al., 2016; Sun, Chen, Liu, Ying, & Xin, 2016; Trendafilov, Fontanella, & Adachi, 2017; Tutz & Schauberger, 2015). However, none of them considered the issue of statistical inference for individual parameters. In principle, the inference methods considered in the current study can be directly applied to all of the above L_1-penalized estimation procedures. For L_1-penalized SEM, the lslx package (Huang, in press) can now run PoSI and PH by specifying the inference argument. It is worth incorporating the reviewed inference methods into other related software implementations as well.

Although postselection inference is now an active research topic in statistics, it is rarely discussed in psychology. Not all existing methods can be extended to SEM in a straightforward manner. Some unadopted yet worthy methods include Tibshirani, Rinaldo, Tibshirani, and Wasserman (2018)'s bootstrap method, Charkhi and Claeskens (2018)'s AIC selector specific method, and Meir and Drton (2017)'s methods using a postselection score function. Studies that tried to improve or extend the reviewed inference methods include a multisplit version of DS (Meinshausen, Meier, & Bühlmann, 2009), a more accurate PH (Liu, Markovic, & Tibshirani, 2018), and PH methods for sequential regression (Tibshirani, Taylor, Lockhart, & Tibshirani, 2016).

Article Information

Conflict of interest disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did report any financial or other conflicts of interest in relation to the work described.

Ethical principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant MOST106-2410-H-006-038 from the Ministry of Science and Technology in Taiwan.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author's institution or the Ministry of Science and

Technology in Taiwan is not intended and should not be inferred.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate: adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81. doi:10.1198/016214504000001907

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid postselection inference. *The Annals of Statistics*, *41*(2), 802–837. doi:10.1214/12-AOS1077

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). Bic and alternative bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 1–19. doi:10.1080/10705511.2014.856691

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. doi:10.1007/BF02294361

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, *87*(419), 738–754. doi:10.1080/01621459.1992.10475276

Brown, L. (1967). The conditional level of student's *t* test. *The Annals of Mathematical Statistics*, *38*(4), 1068–1071. doi:10.1214/aoms/1177698776

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. doi:10.1006/jmps.1999.1279

Buehler, R. J., & Feddersen, A. P. (1963). Note on a conditional property of student's *t*. *The Annals of Mathematical Statistics*, *34*(3), 1098–1100. doi:10.1214/aoms/1177704034

Buja, A., & Zhang, K. (2017). *PoSI: valid postselection inference for linear ls regression [computer software manual]*. R package version 1.0. doi:https://CRAN.R-project.org/package=PoSI

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Bühlmann, P., & van de.Geer, S. (2011). *Statistics for high-dimensional data*. Berlin: Springer.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods & Research*, *21*(1), 89–115. doi:10.1177/0049124192021001004

Casella, G., & Berger, R. (2002). *Statistical inference*. Belmont, CA: Duxbury Press.

Charkhi, A., & Claeskens, G. (2018). Asymptotic postselection inference for the akaike information criterion. *Biometrika*, *105*(3), 645–664. doi:10.1093/biomet/asy018

Chen, P.-Y., & Yao, G. (2015). Measuring quality of life with fuzzy numbers: In the perspectives of reliability, validity, measurement invariance, and feasibility. *Quality of Life Research*, *24*(4), 781–785. doi:10.1007/s11136-014-0816-3

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. doi:10.1080/01621459.2014.934827

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*(1), 115–136. doi:10.1207/s15327906mbr2501_13

Cohen, J. (1992). Uniform asymptotic inference and the bootstrap after model selection. *Psychological Bulletin*, *1(112)*, 155–159. [Mismatch] doi:10.1037/0033-2909.112.1.155

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, *62*(2), 441–444. doi:10.1093/biomet/62.2.441

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*(2), 147–167. doi:10.1207/s15327906mbr1802_2

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, *16*(1), 342–355. doi:10.1214/aos/1176350709

Haughton, D., Oud, J., & Jansen, R. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics—Simulation and Computation*, *26*(4), 1477–1516. doi:10.1080/03610919708813451

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

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*(5), 863–875. doi:10.1007/s11222-014-9458-0

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*(464), 879–899. doi:10.1198/016214503000000828

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. Chicago, IL: University of Chicago Press.

Huang, P.-H. (2017). Asymptotics of aic, bic, and rmsea for model selection in structural equation modeling. *Psychometrika*, *82*(2), 407–426. doi:10.1007/s11336-017-9572-y

Huang, P.-H. (in press). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software*.

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*(2), 329–354. doi:10.1007/s11336-017-9566-9

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Methods*, *14*(1), 6–23. doi:10.1037/a0014694

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structure Equation Modeling*, *23*(4), 555–566. doi:10.1080/10705511.2016.1154793

Jin, S., & Ankargren, S. (2018). Frequentist model averaging in structural equation modelling. *Psychometrika*, *84*, 84–104. doi:10.1007/s11336-018-9624-y

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Thousand Oaks, CA: Sage.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639. doi:10.1080/01621459.1975.10482485

Kabaila, P., & Leeb, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association*, *101*(474), 619–629. doi:10.1198/016214505000001140

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact postselection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927. doi:10.1214/15-AOS1371

Leeb, H., & Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, *19*(1), 100–142. doi:10.1017/S0266466603191050

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59. doi:10.1017/S0266466605050036

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, *34*(5), 2554–2591. doi:10.1214/009053606000000821

Leeb, H., & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(2), 338–376. doi:10.1017/S0266466608080158

Lin, L.-C., Huang, P.-H., & Weng, L.-J. (2017). Selecting path models in sem: A comparison of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 855–869. doi:10.1080/10705511.2017.1363652

Liu, K., Markovic, J., & Tibshirani, R. (2018). More powerful post-selection inference, with application to the Lasso. *ArXiv e-prints*.

Lubke, G. H., & Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 479–490. doi:10.1080/10705511.2016.1141355

Lubke, G. H., Campbell, I., McArtor, D., Miller, P., Luningham, J., & Berg, S. M. V D. (2017). Assessing model selection uncertainty using a bootstrap approach: An update. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 230–245. doi:10.1080/10705511.2016.1252265

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*(1), 107–120. doi:10.1037/0033-2909.100.1.107

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–224. doi:10.1146/annurev.psych.51.1.201

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*(1), 1–32. doi:10.1207/s15327906mbr2901_1

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. doi:10.1037/0033-2909.111.3.490

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). values for high-dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681. p doi:10.1198/jasa.2009.tm08647

Meir, A., & Drton, M. (2017). Tractable postselection maximum likelihood inference for the Lasso. *ArXiv e-Prints*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, *17*(1), 1–14. doi:10.1037/a0026804

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Core Team. doi:https://www.R-project.org/

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Silvey, S. D. (1959). The lagrangian multiplier test. *The Annals of Mathematical Statistics*, *30*(2), 389–407. doi:10.1214/aoms/1177706259

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*(3), 253–263. doi:10.1007/BF02294104

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via l_1 regularization. *Psychometrika*, *81*(4), 921–939. doi:10.1007/s11336-016-9529-6

Taylor, J., & Robert, T. (2018). Postselection inference for l1-penalized likelihood models. *Canadian Journal of Statistics*, *46*(1), 41–61. doi:10.1002/cjs.11313

The WHOQOL Taiwan Group. (2005). *The user's manual of the development of the WHOQOL-BREF taiwan version* (2nd ed.), Taipei: Taiwan WHOQOL Group.

Tibshirani, R. (1996). Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288. doi:10.1111/j.1467-9868.2011.00771.x

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., & Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, *46*(3), 1255–1287. doi:10.1214/17-AOS1584

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact postselection inference for sequential regression procedures. *Journal of the American Statistical*

Association, *111*(514), 600–620. doi:10.1080/01621459.2015.1108848

Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika*, *82*(3), 778–794. doi:10.1007/s11336-017-9575-8

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*(1), 21–43. doi:10.1007/s11336-013-9377-6

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. doi:10.1214/14-AOS1221

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. doi:10.1037/a0027127

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. doi:10.1090/S0002-9947-1943-0012401-3

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. doi:10.2307/1912526

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 397–417. doi:10.1348/000711005X85896

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. doi:10.1214/09-AOS729

Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242. doi:10.1111/rssb.12026

Appendix

Polyhedral method for SEM

The appendix describes how to apply the polyhedral (PH) method to SEM with minimax concave penalty (MCP; Zhang, 2010). Technical details of the PH method for LASSO or for a more general L_1-penalized likelihood can be found in Lee et al. (2016) and Taylor and Robert (2018). The functional form of MCP is

$$\rho\big(|\theta_q|,\lambda\big)=\begin{cases}\lambda|\theta_q|-\dfrac{\theta_q^2}{2\delta} & \text{if } |\theta_q|\le\lambda\delta,\\ \dfrac{1}{2}\lambda^2\delta & \text{if } \lambda\delta<|\theta_q|,\end{cases} \tag{14}$$

where δ is a parameter to control the convexity level of MCP. The L_1 regularization can be seen as a special case of MCP with a convexity parameter of infinity. The penalized likelihood discrepancy with MCP can be written as

$$\mathcal{U}(\theta,\lambda)=\mathcal{D}(\theta)+\mathcal{R}(\phi,\lambda), \tag{15}$$

where $\mathcal{R}(\phi,\lambda)=\sum_{s=1}^{S}\rho(|\phi_s|,\lambda)$ is a penalty term based on MCP. Recall that θ is the model parameter partitioned into (ψ,ϕ), where ψ is freely estimated and ϕ is penalized. Let $\vartheta_{\mathcal{M}}$ denote a subvector of $\theta_{\mathcal{M}}$ formed by $\{\theta_{\mathcal{M},q}\}_{q\in\mathcal{M}}$. It is assumed that $\vartheta_{\mathcal{M}}$ can be written as $\vartheta_{\mathcal{M}}=(\psi_{\mathcal{M}},\varphi_{\mathcal{M}})$. In addition, $\varphi_{\mathcal{M}}^{C}$ can be used to represent the complement of $\varphi_{\mathcal{M}}$, i.e., the vector formed by $\{\theta_{\mathcal{M},q}\}_{q\notin\mathcal{M}}$.

Suppose that $\hat{\theta}_{\mathcal{M}}$ is a minimizer for Equation (15), such that $\hat{\mathcal{M}}=\mathcal{M}$. The minimizer must satisfy the Karush–Kuhn–Tucker (KKT) conditions

$$\begin{aligned}&\frac{\partial\mathcal{D}(\hat{\theta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}}+\frac{\partial\mathcal{R}(\hat{\theta}_{\mathcal{M}},\lambda)}{\partial\vartheta_{\mathcal{M}}}=0,\\ &\left|\frac{\partial\mathcal{D}(\hat{\theta}_{\mathcal{M}})}{\partial\varphi_{\mathcal{M}}^{C}}\right|\le\lambda.\end{aligned} \tag{16}$$

The second inequality implies that each element on the left-hand side should be less than or equal to λ. Note that in the case of L_1 regularization, $\frac{\partial\mathcal{R}(\hat{\theta}_{\mathcal{M}},\lambda)}{\partial\vartheta_{\mathcal{M}}}=(0,\lambda s_{\mathcal{M}})$ with $\hat{s}_{\mathcal{M}}=\text{sign}(\hat{\varphi}_{\mathcal{M}})$.

Let $\bar{\vartheta}_{\mathcal{M}}$ denote the unpenalized ML estimator under $\mathcal{M}$, i.e., $\frac{\partial\mathcal{D}(\bar{\vartheta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}}=0$. For a fixed $\mathcal{M}$, the asymptotic distribution of $\bar{\vartheta}_{\mathcal{M}}$ is

$$\sqrt{N}(\bar{\vartheta}_{\mathcal{M}}-\vartheta_{\mathcal{M}}^{*})\sim\mathcal{N}(0,\hat{\mathcal{C}}_{\mathcal{M}}), \tag{17}$$

where $\hat{\mathcal{C}}_{\mathcal{M}}$ is an estimated covariance matrix of $\bar{\vartheta}_{\mathcal{M}}$ (see Yuan & Hayashi, 2006, for specific formulae). By $\frac{\partial\mathcal{D}(\bar{\vartheta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}}\approx\frac{\partial\mathcal{D}(\hat{\vartheta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}}+\hat{\mathcal{F}}_{\mathcal{M}}(\bar{\vartheta}_{\mathcal{M}}-\hat{\vartheta}_{\mathcal{M}})$, the one-step debiased estimator can be defined as

$$\tilde{\vartheta}_{\mathcal{M}}=\hat{\vartheta}_{\mathcal{M}}-\hat{\mathcal{F}}_{\mathcal{M}}^{-1}\frac{\partial\mathcal{D}(\hat{\theta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}}, \tag{18}$$

where $\hat{\mathcal{F}}_{\mathcal{M}}=\frac{\partial^2\mathcal{D}(\hat{\theta}_{\mathcal{M}})}{\partial\vartheta_{\mathcal{M}}\partial\vartheta_{\mathcal{M}}^{T}}$ is the observed Fisher information matrix under $\mathcal{M}$ (see van de Geer et al., 2014; Zhang & Zhang, 2014). The one-step debiased estimator almost solves the likelihood equation of model $\mathcal{M}$. If λ is chosen, such that $\hat{\vartheta}_{\mathcal{M}}$ is $\sqrt{N}$ consistent with $\vartheta_{\mathcal{M}}^{*}$ (e.g., $\lambda\propto\frac{1}{\sqrt{N}}$), the large sample behavior of $\tilde{\vartheta}_{\mathcal{M}}$ will be quite similar to that of $\bar{\vartheta}_{\mathcal{M}}$. In fact, the asymptotic distribution of $\tilde{\vartheta}_{\mathcal{M}}$ is

$$\sqrt{N}\Big(\tilde{\vartheta}_{\mathcal{M}}-\vartheta_{\mathcal{M}}^{*}\Big)\sim\mathcal{N}(0,\hat{\mathcal{C}}_{\mathcal{M}}), \tag{19}$$

with the constraint

$$\left\{\text{sign}\left(\tilde{\varphi}_{\mathcal{M}}-\mathcal{P}_{\mathcal{M}}\hat{\mathcal{F}}_{\mathcal{M}}^{-1}\frac{\partial\mathcal{R}(\hat{\theta}_{\mathcal{M}},\lambda)}{\partial\vartheta_{\mathcal{M}}}\right)\right\}=\hat{s}_{\mathcal{M}}, \tag{20}$$

where $\hat{s}_{\mathcal{M}}=\text{sign}(\tilde{\varphi}_{\mathcal{M}})$ and $\mathcal{P}_{\mathcal{M}}$ is a projection matrix that selects the rows corresponding to $\tilde{\varphi}_{\mathcal{M}}$ (see Taylor & Robert, 2018). The constraint in Equation (20) can be written in the form of $A\tilde{\vartheta}_{\mathcal{M}}\le b$, i.e.,

$$\begin{pmatrix}0 & -\text{diag}(s_{\mathcal{M}})\end{pmatrix}\tilde{\vartheta}_{\mathcal{M}}\le\begin{pmatrix}0\\ \mathcal{P}_{\mathcal{M}}\hat{\mathcal{F}}_{\mathcal{M}}^{-1}\dfrac{\partial\mathcal{R}(\hat{\theta}_{\mathcal{M}},\lambda)}{\partial\vartheta_{\mathcal{M}}}\end{pmatrix}. \tag{21}$$

In other words, $\tilde{\vartheta}_{\mathcal{M}}$ is asymptotically distributed as a normal variate with mean vector $\vartheta_{\mathcal{M}}^{*}$ and covariance matrix $\frac{1}{N}\hat{\mathcal{C}}_{\mathcal{M}}$ under the polyhedral constraint of $A\tilde{\vartheta}_{\mathcal{M}}\le b$.

Let η denote a vector and $\eta^{T}\vartheta_{\mathcal{M}}^{*}$ be a population target to be inferred. Define $c=\hat{\mathcal{C}}_{\mathcal{M}}\eta(\eta^{T}\hat{\mathcal{C}}_{\mathcal{M}}\eta)^{-1}$ and $r=$

$(I-c\eta^T)\tilde{\vartheta}_{\mathcal{M}}$. The polyhedral lemma (Lee et al., 2016) states that for the random quantity $\eta^T\tilde{\vartheta}_{\mathcal{M}}$ the conditioning set can be written as

$$\left\{A\tilde{\vartheta}_{\mathcal{M}} \le b\right\} = \left\{\mathcal{V}^-(r) \le \eta^T\tilde{\vartheta}_{\mathcal{M}} \le \mathcal{V}^+(r), \mathcal{V}^0(r) \ge 0\right\}, \tag{22}$$

where

$$\begin{aligned} \mathcal{V}^-(r) &= \max_{\{i|(Ac)_i<0\}} \frac{b_i-(Ar)_i}{(Ac)_i} \\ \mathcal{V}^+(r) &= \min_{\{i|(Ac)_i>0\}} \frac{b_i-(Ar)_i}{(Ac)_i} \\ \mathcal{V}^0(r) &= \min_{\{i|(Ac)_i=0\}} b_i-(Ay)_i. \end{aligned} \tag{23}$$

Note that $\eta^T\tilde{\vartheta}_{\mathcal{M}}$ and $(\mathcal{V}^-(r), \mathcal{V}^+(r), \mathcal{V}^0(r))$ are statistically independent. The polyhedral lemma further says that the large sample distribution of $\eta^T\tilde{\vartheta}_{\mathcal{M}}$ given $A\tilde{\vartheta}_{\mathcal{M}} \le b$ and r is

$$\eta^T\tilde{\vartheta}_{\mathcal{M}} | \left\{A\tilde{\vartheta}_{\mathcal{M}} \le b, r\right\} \sim F^{\mathcal{V}^-(r),\mathcal{V}^+(r)}_{\eta^T\vartheta^*_{\mathcal{M}},\frac{1}{N}\eta^T\hat{\mathcal{C}}_{\mathcal{M}}\eta}(t), \tag{24}$$

i.e., a truncated normal with mean $\eta^T\vartheta^*_{\mathcal{M}}$, variance $\frac{1}{N}\eta^T\hat{\mathcal{C}}_{\mathcal{M}}\eta$, lower limit $\mathcal{V}^-(r)$, and upper limit $\mathcal{V}^+(r)$. Here, $F^{L,U}_{\mu,\sigma^2}(t)$ denotes the cumulative distribution function of a truncated normal variate T. Therefore, we can construct a $1-\alpha$ CI for $\vartheta^*_{\mathcal{M},q} = \eta^T\vartheta^*_{\mathcal{M}}$ by setting $C^{PH}_{\mathcal{M},q} = [L, U]$, such that

$$F^{\mathcal{V}^-(r),\mathcal{V}^+(r)}_{L,\frac{1}{N}\eta^T\hat{\mathcal{C}}_{\mathcal{M}}\eta}\left(\eta^T\hat{\vartheta}_{\mathcal{M}}\right) = 1-\frac{\alpha}{2}, \tag{25}$$

and

$$F^{\mathcal{V}^-(r),\mathcal{V}^+(r)}_{U,\frac{1}{N}\eta^T\hat{\mathcal{C}}_{\mathcal{M}}\eta}\left(\eta^T\hat{\vartheta}_{\mathcal{M}}\right) = \frac{\alpha}{2}. \tag{26}$$

It should be noted that the constructed PH interval is actually conditioned on the selection event $\hat{\mathcal{M}} = \mathcal{M}$ and $\hat{s} = s_{\mathcal{M}}$, where $s_{\mathcal{M}}$ is the sign of $\hat{\vartheta}_{\mathcal{M}}$. If we wish to condition on $\hat{\mathcal{M}} = \mathcal{M}$, Lee et al. (2016) suggests using $\cup_s[\mathcal{V}^-_s(r), \mathcal{V}^+_s(r)]$, the union of $[\mathcal{V}^-_s(r), \mathcal{V}^+_s(r)]$ among all possible signs of $\mathcal{M}$. However, this union is difficult to evaluate when the number of penalized parameters is large. The current study simply calculated PH intervals conditioned on both $\hat{\mathcal{M}} = \mathcal{M}$ and $\hat{s} = s_{\mathcal{M}}$, which generally yields wider intervals (see Lee et al., 2016).