PSYCHOMETRIKA—VOL. 82, NO. 2, 407–426
JUNE 2017
DOI: 10.1007/s11336-017-9572-y

ASYMPTOTICS OF AIC, BIC, AND RMSEA FOR MODEL SELECTION IN STRUCTURAL EQUATION MODELING

PO-HSIEN HUANG

NATIONAL CHENG KUNG UNIVERSITY

Model selection is a popular strategy in structural equation modeling (SEM). To select an "optimal" model, many selection criteria have been proposed. In this study, we derive the asymptotics of several popular selection procedures in SEM, including AIC, BIC, the RMSEA, and a two-stage rule for the RMSEA (RMSEA-2S). All of the results are derived under weak distributional assumptions and can be applied to a wide class of discrepancy functions. The results show that both AIC and BIC asymptotically select a model with the smallest population minimum discrepancy function (MDF) value regardless of nested or non-nested selection, but only BIC could consistently choose the most parsimonious one under nested model selection. When there are many non-nested models attaining the smallest MDF value, the consistency of BIC for the most parsimonious one fails. On the other hand, the RMSEA asymptotically selects a model that attains the smallest population RMSEA value, and the RESEA-2S chooses the most parsimonious model from all models with the population RMSEA smaller than the pre-specified cutoff. The empirical behavior of the considered criteria is also illustrated via four numerical examples.

Key words: structural equation modeling, Akaike information criterion, Bayesian information criterion, root-mean-square error of approximation, model selection.

1. Introduction

Model comparison (or alternative models) is one of the main strategies for conducting structural equation modeling (SEM; Jöreskog, 1993; MacCallum & Austin, 2000). When utilizing this strategy, several candidate models are formulated, and then an optimal one is chosen from them based on some decision rule. The candidate models are often specified to represent different psychological theories to explain the covariance matrix among variables (e.g., Keyes, Shmotkin, & Ryff, 2002). The optimality of a model is often defined through the model goodness of fit and the model complexity[1] (or parsimony; Pitt, Myung, & Zhang, 2002; Preacher, 2006). Some methodologists advocate the application of model comparisons because the relative advantages and disadvantages of several substantive theories can be compared in a single study (e.g., Burnham & Anderson, 2002; MacCallum, 2003; MacCallum & Austin, 2000). A review by MacCallum and Austin (2000) also showed that about 50% of SEM applications utilize model comparison strategies to answer research questions.

The process of selecting an optimal model from a set of candidate models is called model selection in statistical literature. In practice, the model selection task is usually achieved through optimizing the value of a specific model selection criterion. A lot of selection criteria have been

The research was supported in part by Grant MOST 104-2410-H-006-119-MY2 from the Ministry of Science and Technology in Taiwan. The author would like to thank Wen-Hsin Hu and Tzu-Yao Lin for their help in simulating data.

Electronic supplementary material The online version of this article (doi:10.1007/s11336-017-9572-y) contains supplementary material, which is available to authorized users.

Correspondence should be made to Po-Hsien Huang, Department of Psychology, National Cheng Kung University, No.1, University Road, Tainan City, 701, Taiwan. Email: psyphh@mail.ncku.edu.tw

[1] Both model goodness of fit and model complexity (or parsimony) are broad concepts, and researchers may interpret them in different ways. In the current study, model goodness of fit is measured by some minimum discrepancy function as introduced in Section 2, and model complexity is simply represented by the number of parameters. For readers who are interest in further discussion on model goodness of fit and model complexity, please refer to Preacher (2006).

proposed historically, including Akaike information criterion (AIC; Akaike, 1974), Mallow's $\mathcal{C}_p$ (Mallow, 1973), delete-one cross-validation (Stone, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and generalized cross-validation (Wahba, 1990). Shao (1997) provided an excellent review of model selection criteria in the context of a linear regression analysis. Among these criteria, AIC and BIC are the two most well known. Based on the derivation of AIC and BIC, the former aims to choose a model with minimal Kullback–Leibler divergence (Kullback & Leibler, 1951), and the latter is targeted toward the selection of a model with the maximum posterior probability given the data. In a linear regression analysis, the asymptotic behavior of AIC and BIC is well understood (see Shao, 1997 for a review). The general results can be summarized as follows: When the true model is infinitely dimensional, AIC is asymptotically loss efficient in the sense that it selects a model with nearly minimum risk; on the other hand, when the candidate models contain a true model with a finite dimension, BIC can select the true model consistently.

In the context of SEM, some theoretical results for AIC and BIC have also been derived. Bozdogan (1987) pointed out the inconsistency of AIC in selecting the true model and proposed a consistent version of AIC, called consistent ACI (CAIC). Haughton, Oud, and Jansen (1997) also studied the consistency issue of AIC and BIC with heuristic arguments and conducted simulations to support their theoretical results. However, the arguments in Bozdogan as well as Haughton et al. both rely on the Chi-square approximation of log-likelihood ratio statistics. In real SEM applications, the Chi-square approximation generally fails due to the violation of the normality assumption (Micceri, 1989) and the misspecification of candidate models (Cudeck & Henly, 1991; MacCallum, 2003). Hence, it is particularly of interest to understand the asymptotics of AIC and BIC under non-normality and misspecification of all the candidate models. The first goal of this study is to answer this question rigorously. The results show that even when the data is non-normal, and all the candidate models are wrong, BIC is still consistent for some quasi-true models, but AIC is only consistent for the models with smallest minimum discrepancy function (MDF) values. The so-called quasi-true model here is defined through the population MDF value and the number of freely estimated parameters (see $\mathcal{A}_d^*$ in Equation (9)).

Two types of model selection can be distinguished: nested model selection and non-nested model selection. The relation between two models is said to be nested if one of them can be seen as a special case of the other by adding constraints on the parameters; otherwise, the relation is said to be non-nested. Previous theoretical results were established for the candidate models with nested relations. In SEM practice, however, alternative models based on different theoretical grounds are sometimes non-nested. Because AIC and BIC are often suggested for non-nested model selection (e.g., Jöreskog, 1993; Kaplan, 2009; West, Taylor, & Wu, 2012), the second goal of the present research is to study the limiting behaviors of AIC and BIC for the case of non-nested model selection.

In SEM, the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980) is a popular goodness-of-fit index that measures the misfit of a specified model per degree of freedom. Unlike other goodness-of-fit indices, the RMSEA can be used in both a descriptive and inferential manner (e.g., Browne & Cudeck, 1993; Li & Bentler, 2006). Recently, the RMSEA is also being treated as a model selection criterion, and simulation results show that it outperforms AIC and BIC with regard to selecting an approximately correct model (Preacher, Zhang, Kim, & Mels, 2013). Therefore, the third goal of the study is to derive the asymptotics of the RMSEA for model selection.

2. Notations and Settings

Let Y denote a P-dimensional random vector from a distribution F with a zero mean and covariance Σ. Given a centered random sample, $\mathcal{Y}_N = \{Y_n\}_{n=1}^N$, a consistent estimator of Σ

can be obtained through $S_N = \frac{1}{N}\sum_{n=1}^{N} Y_n Y_n^T$. We use $\sigma = \text{vech}(\Sigma)$ and $s = \text{vech}(S)$ to denote the vectors that contain the P^* non-duplicated elements of Σ and S, respectively, where $P^* = P(P+1)/2$.

Definition 1. An SEM model $\Sigma_\alpha(\theta_\alpha)$ indexed by $\alpha \in \mathcal{A}$ is a function from Θ_α to $\mathcal{S}_{++}^P$, where $\mathcal{A}$ is an index set; $\Theta_\alpha \subset \mathbb{R}^{|\alpha|}$ is the parameter space of θ_α; $\mathcal{S}_{++}^P$ is the set formed by all symmetric positive definite matrix in $\mathbb{R}^{P\times P}$, and $|\alpha|$ is the dimension of θ_α.

Because $\Sigma_\alpha(\theta_\alpha)$ is symmetrical, we can also use $\sigma_\alpha(\theta_\alpha) = \text{vech}(\Sigma_\alpha(\theta_\alpha))$ to represent an SEM model. $\sigma_\alpha(\theta_\alpha)$ is now a function from Θ_α to $\mathcal{M}^{P^*} \subset \mathbb{R}^{P^*}$, where $\mathcal{M}^{P^*}$ is the range of $\mathcal{S}_{++}^P$ under the transformation vech$(\cdot)$.

Definition 2. A discrepancy function $\mathcal{D}(\cdot,\cdot)$ is defined as a function from $\mathcal{M}^{P^*} \times \mathcal{M}^{P^*}$ to $\mathbb{R}^+ \cup \{0\}$ such that $\mathcal{D}(\sigma_\alpha(\theta_\alpha), \sigma) = 0$ if and only if $\sigma_\alpha(\theta_\alpha) = \sigma$.

In practice, the most commonly used discrepancy function is the maximum likelihood (ML) fitting function (see Jackson, Gillaspy, & Pure-Stephenson, 2009; Shah & Goldstein, 2006 for reviews)

$$\mathcal{D}_{ML}(\sigma_\alpha(\theta_\alpha), \sigma) = -\log\left|\Sigma_\alpha(\theta_\alpha)^{-1}\Sigma\right| + \text{tr}\left(\Sigma_\alpha(\theta_\alpha)^{-1}\Sigma\right) - P. \tag{1}$$

Other well-known discrepancy functions include ordinary least squares (OLS), weighted least squares (WLS; Browne, 1974), and generalized least squares (GLS; Browne, 1984). The family of least squares discrepancy function can be represented as

$$\mathcal{D}_{LS}(\sigma_\alpha(\theta_\alpha), \sigma) = (\sigma_\alpha(\theta_\alpha) - \sigma)^T W (\sigma_\alpha(\theta_\alpha) - \sigma), \tag{2}$$

where W is a $P^* \times P^*$ positive definite weight matrix. In practice, W is usually replaced by its estimator $\hat{W}_N$. As a result, the asymptotic property of LS estimation relies on the consistency of $\hat{W}_N$ for W. In subsequent discussion, the consistency of $\hat{W}_N$ for W is always assumed.

Definition 3. Given a true covariance vector σ^0, an SEM model $\sigma_\alpha(\theta_\alpha)$ is said to be correct for σ^0 if there exists a parameter value θ_α^0 such that $\mathcal{D}\left(\sigma_\alpha\left(\theta_\alpha^0\right), \sigma^0\right) = 0$; otherwise, $\sigma_\alpha(\theta_\alpha)$ is said to be incorrect for σ^0. When $\sigma_\alpha(\theta_\alpha)$ is incorrect, a quasi-true parameter value θ_α^* is defined as a minimizer of $\mathcal{D}\left(\sigma_\alpha(\theta_\alpha), \sigma^0\right)$.

By the property of discrepancy function, θ_α^0 can be seen as a special case of θ_α^*, with $\mathcal{D}\left(\sigma_\alpha\left(\theta_\alpha^0\right), \sigma^0\right) = 0$. Hence, we will use θ_α^* to represent both the quasi-true and true parameter values in subsequent discussion.

Definition 4. Given a model $\sigma_\alpha(\theta_\alpha)$ and a sample covariance vector s, a minimum discrepancy function (MDF) estimator $\hat{\theta}_\alpha$ (with respect to $\mathcal{D}$) is defined as a minimizer of $\mathcal{D}(\sigma_\alpha(\theta_\alpha), s)$.

In later discussion, we simply use $\sigma(\alpha)$, $\hat{\sigma}(\alpha)$, and $\sigma^*(\alpha)$ to denote $\sigma_\alpha(\theta_\alpha)$, $\sigma_\alpha\left(\hat{\theta}_\alpha\right)$, and $\sigma_\alpha\left(\theta_\alpha^*\right)$, respectively. Note that $\sigma_\alpha\left(\hat{\theta}_\alpha\right)$ and $\sigma_\alpha\left(\theta_\alpha^*\right)$ are quantities depending on the values of s and σ^0, respectively, although we omit that dependency in their notations. Similarly, $\hat{\mathcal{D}}(\alpha) \equiv \mathcal{D}\left(\hat{\sigma}(\alpha), s\right)$ and $\mathcal{D}^*(\alpha) \equiv \mathcal{D}\left(\sigma^*(\alpha), \sigma^0\right)$ are used to represent the estimated and population MDF value under model α.

Definition 5. Given a set of candidate models $\mathcal{A}$, a model selection procedure is a decision rule that chooses an “optimal” model $\hat{\alpha}_N$ from $\mathcal{A}$ based on a specific model selection criterion $\mathcal{C}(\alpha, \mathcal{D}, s)$.

The subscript N of $\hat{\alpha}_N$ is used to emphasize the dependence of $\hat{\alpha}_N$ on a random sample. In this study, we consider the following three model selection criteria:

$$AIC_N(\alpha) \equiv \hat{\mathcal{D}}(\alpha) + \frac{2}{N}|\alpha|, \tag{3}$$

$$BIC_N(\alpha) \equiv \hat{\mathcal{D}}(\alpha) + \frac{\log N}{N}|\alpha|, \tag{4}$$

and

$$RMSEA_N(\alpha) \equiv \sqrt{\max\left\{\frac{\hat{\mathcal{D}}(\alpha)}{df(\alpha)} - \frac{1}{N}, 0\right\}}, \tag{5}$$

where $df(\alpha) = P^* - |\alpha|$. In the simplest way, model selection procedures based on $AIC_N(\alpha)$, $BIC_N(\alpha)$, or $RMSEA_N(\alpha)$ select a model that attains the minimum value of the corresponding criterion. Strictly speaking, the criteria in Equations (3) and (4) are not really the AIC and the BIC proposed by Akaike (1974) and Schwarz (1978) because the two indices are proposed under the ML framework, which implies that $\hat{\mathcal{D}}(\alpha)$ should be $\hat{\mathcal{D}}_{ML}(\alpha)$. However, in Section 3, we will show that the selection criterion in the form of either Equation (3) or (4) has the same asymptotic behavior, regardless of which discrepancy function is used.

Preacher, Zhang, Kim, and Mels (2013) suggested another two-stage decision rule for RMSEA (RMSEA-2S). In the first stage, models with $RMSEA_N(\alpha) \leq c$ are collected, where c is a pre-specified nonnegative cutoff. Usually, c is set as .05 based on the recommendation of Browne and Cudeck (1993). In the second stage, the model with smallest number of parameters is chosen from the output of the first stage. Therefore, the two-stage rule is to choose the most parsimonious model among all the models that fit the data reasonably well in terms of RMSEA.

Given a set of candidate models $\mathcal{A}$, we partition $\mathcal{A}$ into $\mathcal{A}_- = \{\alpha | \mathcal{D}^*(\alpha) > 0\}$ and $\mathcal{A}_+ = \{\alpha | \mathcal{D}^*(\alpha) = 0\}$. $\mathcal{A}_-$ and $\mathcal{A}_+$ contains all the incorrect and correct models from $\mathcal{A}$, respectively. Either $\mathcal{A}_-$ or $\mathcal{A}_+$ can be empty. However, because psychological theories cannot perfectly explain human behavior, we may expect that $\mathcal{A}_+ = \emptyset$ in practice.

One possible method for comparing the appropriateness of the candidate models is to compare their MDF values. Hence, the first optimal set of models can be defined as

$$\mathcal{A}_d = \left\{\alpha | \mathcal{D}^*(\alpha) = \min_{\alpha' \in \mathcal{A}} \mathcal{D}^*\left(\alpha'\right)\right\}. \tag{6}$$

The "d" stands for "discrepancy" because $\mathcal{A}_d$ contains all the candidate models with the smallest MDF value. Another way is to compare their MDF values divided by the corresponding degrees of freedom, the population RMSEA values. Based on this idea, the second optimal set is defined as

$$\mathcal{A}_e = \left\{\alpha | \mathcal{D}^*(\alpha)/df(\alpha) = \min_{\alpha' \in \mathcal{A}} \mathcal{D}^*\left(\alpha'\right)/\mathrm{d}f\left(\alpha'\right)\right\}. \tag{7}$$

The "e" denotes "effectiveness" because $\mathcal{A}_e$ emphasizes the effectiveness of each parameter to explain the covariance. Based on the decision rule of Preacher et al. (2013), we also consider the following optimal set for RMSEA

$$\mathcal{A}_c = \left\{\alpha | \mathcal{D}^*(\alpha)/df(\alpha) \leq c\right\}, \tag{8}$$

where "c" stands for "cutoff". $\mathcal{A}_c$ collects all the models with a population RMSEA smaller than or equal to c. Unlike $\mathcal{A}_d$ and $\mathcal{A}_e$, $\mathcal{A}_c$ can be empty if there is no candidate model satisfying $\mathcal{D}^*(\alpha)/df(\alpha) \leq c$. When $\mathcal{A}_d$, $\mathcal{A}_e$, or $\mathcal{A}_c$ contains more than one model, we may prefer more parsimonious models in $\mathcal{A}_d$, $\mathcal{A}_e$, or $\mathcal{A}_c$. Therefore, three subsets of $\mathcal{A}_d$, $\mathcal{A}_e$, and $\mathcal{A}_c$ are further defined as

$$\mathcal{A}_d^* = \left\{\alpha | \, \alpha \in \mathcal{A}_d, |\alpha| = \min_{\alpha' \in \mathcal{A}_d} \left|\alpha'\right|\right\}, \tag{9}$$

$$\mathcal{A}_e^* = \left\{\alpha | \, \alpha \in \mathcal{A}_e, |\alpha| = \min_{\alpha' \in \mathcal{A}_e} \left|\alpha'\right|\right\}, \tag{10}$$

and

$$\mathcal{A}_c^* = \left\{\alpha | \, \alpha \in \mathcal{A}_c, |\alpha| = \min_{\alpha' \in \mathcal{A}_c} \left|\alpha'\right|\right\}. \tag{11}$$

$\mathcal{A}_d^*$, $\mathcal{A}_e^*$, and $\mathcal{A}_c^*$ collect the models with a minimum number of parameters in $\mathcal{A}_d$, $\mathcal{A}_e$, and $\mathcal{A}_c$, respectively. The models in $\mathcal{A}_d^*$ are the so-called quasi-true model mentioned in Section 1. $\mathcal{A}_e^*$ and $\mathcal{A}_c^*$ are new and can be used to describe the asymptotic behavior of the RMSEA. If $\mathcal{A}_+ \neq \emptyset$, we have $\mathcal{A}_d^* = \mathcal{A}_e^* \subset \mathcal{A}_d = \mathcal{A}_e = \mathcal{A}_+ \subset \mathcal{A}_c$ and $\mathcal{A}_c^* \subset \mathcal{A}_c$; otherwise, we only have $\mathcal{A}_d^* \subset \mathcal{A}_d$, $\mathcal{A}_e^* \subset \mathcal{A}_e$, and $\mathcal{A}_c^* \subset \mathcal{A}_c$.

In later discussion, we assume that $\mathcal{A}_d^*$ and $\mathcal{A}_e^*$ are both singletons to simplify our problem, i.e., $\mathcal{A}_d^* = \left\{\alpha_d^*\right\}$ and $\mathcal{A}_e^* = \left\{\alpha_e^*\right\}$. Note that the assumption is not unreasonable. When $\mathcal{A}$ is formed by a series of nested models, the assumption must be true. Under non-nested settings, the violation of this assumption means that there exist at least two models, denoted by α_1^* an α_2^*, such that they attain the minimal values of MDF or RMSEA in population with the same model complexity that differ in terms of their functional forms. The existence of such α_1^* and α_2^* is conceptually possible. However, we think that it would be extremely rare to encounter such cases in actual research settings. Even if we can deliberately construct candidate models in simulations, it is still difficult to obtain this type of α_1^* and α_2^*.

Definition 6. A model selection procedure is said to be consistent for $\mathcal{A}^* \subset \mathcal{A}$ if

$$\mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}^*\right) \to 1, \tag{12}$$

as $N \to \infty$. In particular, if $\mathcal{A}^* = \{\alpha^*\}$, i.e., $\mathcal{A}^*$ is a singleton, we say the procedure is consistent for α^*.

Clearly, the consistency of a model selection procedure for some $\mathcal{A}^* \subset \mathcal{A}$ is a crucial property if we hope to understand its asymptotic behavior. Note that our definition of consistency is different from that of Shao (1997). In Shao's work, a selection procedure is consistent if it can always choose a model α that minimizes a sample-dependent loss. In SEM settings, the sample-dependent loss is $\mathcal{D}\left(\hat{\sigma}(\alpha), \sigma\right)$. On the other hand, our definition relies on some optimal set $\mathcal{A}^*$ determined by the population MDF value $\mathcal{D}\left(\sigma^*(\alpha), \sigma^0\right)$.

In later discussion, we assume that the following regularity conditions hold.

Condition A. $\mathcal{A}$ is pre-specified, with $|\mathcal{A}| = K < \infty$, and $\mathcal{A}_d^* = \left\{\alpha_d^*\right\}$ and $\mathcal{A}_e^* = \left\{\alpha_e^*\right\}$ are both singletons.

Condition B. $\sqrt{N}\left(s - \sigma^0\right) \longrightarrow_L N(0, \Gamma)$, and there exists an estimator $\hat{\Gamma}$ such that $\hat{\Gamma} \longrightarrow_P \Gamma$, where $\longrightarrow_L$ denotes "converge in law," and $\longrightarrow_P$ denotes "converge in probability."

Condition C. For each $\alpha \in \mathcal{A}$, $\sigma_\alpha(\theta_\alpha)$ is continuously twice differentiable.

Condition D. For each $\alpha \in \mathcal{A}$, $\mathcal{D}(\sigma_\alpha(\theta_\alpha), \sigma)$ is continuously twice differentiable in both arguments.

Condition E. For each $\alpha \in \mathcal{A}$ and for all $\sigma^0 \in \mathcal{M}^{P^*}$, there exist a $\theta_\alpha^* \in \Theta_\alpha$ such that (1) θ_α^* is the unique minimizer of $\mathcal{D}\left(\sigma_\alpha(\theta_\alpha), \sigma^0\right)$; (2) θ_α^* is an interior point of a compact parameter set Θ_α; (3) θ_α^* is a regular point of $\frac{\partial^2 \mathcal{D}(\sigma_\alpha(\theta_\alpha),\sigma)}{\partial\theta_\alpha \partial\theta_\alpha^T}$ with rank $|\alpha|$ for all σ in the neighborhood of σ^0.

Condition A requires that the candidate set contains only finite pre-specified models. When all the candidate models are formulated in advance, the condition is of course satisfied. Even in a very exploratory setting, if the researcher can consider all possible types of alternatives a priori, the condition will still be satisfied. Condition B is a standard assumption for SEM. If each observed variable has a finite $4+\delta$ moment for some $\delta > 0$, Condition B will be true. Condition C assumes that each model is smooth enough. If the specified model is in the class of Bentler and Weeks (1980), Condition C is generally true. Condition D implies that the discrepancy function can be approximated by a quadratic function in the neighborhood of some chosen point. Condition E describes the existence, uniqueness, and the geometry of quasi-true parameter θ_α^*. Conditions B, C, D, and E are sufficient for the consistency of an MDF estimator, i.e., $\hat{\theta}_\alpha \longrightarrow_P \theta_\alpha^*$ (Shapiro, 1984, Theorem 1), and its asymptotic normality (Shapiro, 1983, Theorem 5.4).

3. Main Results

In this section, four theorems are derived to describe the large sample behavior of AIC, BIC, RMSEA, and RMSEA-2S. All of the proofs are given in Appendix. Since AIC, BIC, and many other information criteria can be written in the form of $\hat{\mathcal{D}}(\alpha) + k_N |\alpha|$ for some deterministic or stochastic sequence k_N. In later discussion, we use $IC_{k_N}(\alpha)$ to represent $\hat{\mathcal{D}}(\alpha) + k_N |\alpha|$ for simplicity and derive the asymptotic properties of $IC_{k_N}(\alpha)$ under different orders of k_N.

Theorem 1. *Let $\hat{\alpha}_N$ denote the model selection result by minimizing $IC_{k_N}(\alpha)$ for $k_N = O_{\mathbb{P}}\left(N^{-1}\right)$. Then*

(1) $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d\right) = 1$;
(2) If $\mathcal{A}_d \backslash \alpha_d^* \neq \emptyset$, $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) > 0$.

Theorem 1 describes the limiting behavior of any IC_{k_N} with $k_N = O_{\mathbb{P}}\left(N^{-1}\right)$, where $O_{\mathbb{P}}(\cdot)$ denotes the stochastic big O notation. AIC is obviously a special case of this form, with $k_N = \frac{2}{N}$. Part (1) of Theorem 1 shows that AIC asymptotically selects a model belonging to $\mathcal{A}_d$. Hence, $\hat{\mathcal{D}}\left(\hat{\alpha}_N\right)$ asymptotically attains the smallest $\mathcal{D}^*(\alpha)$ on $\mathcal{A}$. However, if $\mathcal{A}_d$ contains a model with unnecessary parameters, part (2) of Theorem 1 indicates that AIC may choose a model with unnecessary parameters; i.e., AIC is generally not consistent for α_d^*. Of course, if $\mathcal{A}_d = \left\{\alpha_d^*\right\}$, AIC is consistent for α_d^*.

Remark 1. Another well-known criterion belonging to this class is AIC3, which utilizes $k_N = \frac{3}{N}$ (Sclove, 1987). Although AIC and AIC3 have the same large sample properties, their finite sample behavior can be different. For example, AIC3 has been shown to outperform AIC in selecting the correct numbers of factors (Dziak et al., 2012).

Theorem 2. *Let $\hat{\alpha}_N$ denote the model selection result by minimizing IC_{k_N} with k_N satisfying $\sqrt{N} k_N = O_{\mathbb{P}}(1)$ and $N k_N \longrightarrow_P +\infty$. Then*

(1) $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d\right) = 1$;
(2) If $\mathcal{A}_d \backslash \alpha_d^* \neq \emptyset$, and $\sigma^*\left(\alpha_d^*\right) = \sigma^*(\alpha)$ for each $\alpha \in \mathcal{A}_d \backslash \alpha_d^*$, then we have $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N = \alpha_d^*\right) = 1$;
(3) If $\mathcal{A}_d \backslash \alpha_d^* \neq \emptyset$, and $\sigma^*\left(\alpha_d^*\right) \neq \sigma^*(\alpha)$ for some $\alpha \in \mathcal{A}_d \backslash \alpha_d^*$, then $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) > 0$.

Theorem 2 can be used to describe the large sample behavior of BIC since $k_N = \frac{\log N}{N}$ satisfies $\sqrt{N} k_N = O_{\mathbb{P}}(1)$ and $N k_N \longrightarrow_P +\infty$. Again, Part (1) of Theorem 2 shows that BIC asymptotically selects a model belonging to $\mathcal{A}_d$. Part (2) of Theorem 2 further shows that BIC is consistent for α_d^* when the population model-implied covariance of each model in $\mathcal{A}_d \backslash \alpha_d^*$ is equal to that of α_d^*. If $\mathcal{A}$ is formed by a series of nested models, we can expect that $\sigma^*\left(\alpha_d^*\right) = \sigma^*(\alpha)$ for any $\alpha \in \mathcal{A}_d \backslash \alpha_d^*$ and hence that BIC is consistent for α_d^*. However, for a case where some pair of α_d^* and $\alpha \in \mathcal{A}_d \backslash \alpha_d^*$ is non-nested, the consistency of BIC for $\mathcal{A}_d^*$ may fail, as indicated by Part (3). A simple way to handle the inconsistency of BIC under general non-nested selection is to consider a heavier penalty $k_N |\alpha|$ satisfying $k_N = o_{\mathbb{P}}(1)$ and $\sqrt{N} k_N \longrightarrow_P +\infty$, such as $\frac{\log N}{\sqrt{N}} |\alpha|$. However, we do not recommend using this penalty in practice because $\frac{\log N}{\sqrt{N}}$ is too heavy, which may result in poor finite sample performance.

Remark 2. Besides BIC, many selection criteria can be also written as IC_{k_N}, with k_N satisfying $\sqrt{N} k_N = O_{\mathbb{P}}(1)$ and $N k_N \longrightarrow_P +\infty$. These criteria include consistent AIC (Bozdogan, 1987), the Hannan–Quinn information criterion (Hannan & Quinn, 1979), Haughton's BIC (Haughton, 1988), and the sample adjusted BIC (Sclove, 1987). Theorem 2 can be applied to all of these criteria.

Theorems 1 and 2 show that both AIC and BIC asymptotically select a model with the smallest MDF value on the candidate set, which is consistent with previous simulation results (e.g., Bollen, Ray, Zavisca, & Harden, 2014; Haughton, Oud, & Jansen, 1997; Homburg, 1991). Interestingly, the limiting model selected by either AIC or BIC is not really a compromise in terms of models goodness of fit and complexity. Goodness of fit has a priority role in defining the limiting model chosen by either AIC or BIC (see the definition of $\mathcal{A}_d$ and $\mathcal{A}_d^*$). In a nested model selection setting, if each added parameter actually improves the model fit, both AIC and BIC ultimately select the most complex model. The classical example of Cudeck and Henly (1991) showed this phenomenon. This fact can also explain the simulation results of Preacher, Zhang, Kim, and Mels (2013), who found that AIC and BIC cannot consistently select the researcher-defined approximately true model and tend to choose a model with a lower MDF value.

Theorem 3. *Let $\hat{\alpha}_N$ denote the model selection result by minimizing $RMSEA_N(\alpha)$. Then*

(1) $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_e\right) = 1$;
(2) If $\mathcal{A}_e \backslash \alpha_e^* \neq \emptyset$, then $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_e \backslash \alpha_e^*\right) > 0$.

Part (1) of Theorem 3 shows that by minimizing the RMSEA, $\hat{\alpha}_N$ is consistent for $\mathcal{A}_e$. Because $\mathcal{A}_e$ and $\mathcal{A}_d$ are generally not equal under $\mathcal{A}_+ = \emptyset$, a model selected by the RMSEA can be quite different from one selected by either AIC or BIC. If $\mathcal{A}_e \backslash \alpha_e^*$ is not empty, the RMSEA may select a model in $\mathcal{A}_e \backslash \alpha_e^*$, as indicated by Part (2) of Theorem 3. Compared to the limiting behaviors of AIC and BIC, the RMSEA can select a model that simultaneously takes into account both model fit and model complexity.

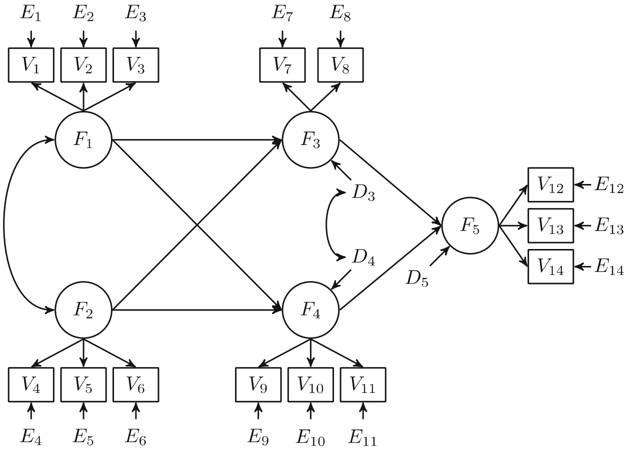

FIGURE 1.
Model specification for model α_4.

Theorem 4. *Let $\hat{\alpha}_N$ denote the model selection result based on the two-stage decision rule for the RMSEA. Then*

(1) If $\mathcal{A}_c = \emptyset$, then the procedure selects nothing asymptotically;
(2) If $\mathcal{A}_c \neq \emptyset$ and $\mathcal{D}^* (\alpha_c) / df (\alpha_c) < c$ for all $\alpha_c \in \mathcal{A}_c$, then $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_c^*\right) = 1$.

Theorem 4 describe the asymptotic behavior of the RMSEA-2S. Part (1) is interesting because it shows that we can reject all of the models if no candidate is good enough. Part (2) shows that when all the models in $\mathcal{A}_c$ have $\mathcal{D}^* (\alpha_c) / df (\alpha_c) < c$, the two-stage rule is consistent for $\mathcal{A}_c^*$. If some α_c has $\mathcal{D}^* (\alpha_c) / df (\alpha_c) = c$, it will not be consistently selected in the first stage. The model chosen by the RMSEA-2S is generally not the model with the smallest $\mathcal{D}^* (\alpha) / df (\alpha)$, which can be used to explain why this rule can select a researcher-defined approximately true model, as shown in Preacher et al. (2013).

Remark 3. Preacher et al. (2013) also suggested a modified RMSEA-2S. Under this modified procedure, the models with $LB_N (\alpha) \leq c$ are collected in the first stage, where $LB_N (\alpha)$ is the 95% lower limit for $\mathcal{D}^* (\alpha) / df (\alpha)$. Since $LB_N (\alpha)$ converges to $RMSEA_N (\alpha)$ from below, this modified rule is also consistent for $\mathcal{A}_c^*$.

4. Numerical Illustrations

In this section, the empirical behaviors of AIC, BIC, the RMSEA, and the RMSEA-2S under four different settings are illustrated. In Setting A, the MDF value of each candidate model is constructed to be different. In settings B and C, more than one candidate model attains the smallest MDF value; i.e., $\mathcal{A}_d$ is not a singleton, but these models have different numbers of parameters, which implies that $\mathcal{A}_d \backslash \alpha_d^* \neq \emptyset$. The main difference between these two settings is that under Setting B, models in $\mathcal{A}_d$ are all correct, but under Setting C, none of these models are correct. Setting D is the most complex: No candidate models are correct; $\mathcal{A}_d$ and $\mathcal{A}_e$ are different, and neither $\mathcal{A}_d \backslash \alpha_d^*$ and $\mathcal{A}_e \backslash \alpha_e^*$ are empty.

The data sets in this numerical illustration are based on the sample covariance matrix in a study by McDonald (McDonald, 2010; see also Feist, Bodner, Jacobs, Miles, & Tan, 1995).

TABLE 1.
The details of model construction of $\alpha_1 - \alpha_3$ and $\alpha_5 - \alpha_7$ based on specification of α_4.

Model	α_1	α_2	α_3	α_5	α_6	α_7
Construction	Delete $F_1 \rightarrow F_3$ $F_2 \rightarrow F_4$	Delete $F_1 \rightarrow F_3$	Delete $D_3 \leftrightarrow D_4$	Add $E_7 \leftrightarrow E_9$	Add $E_8 \leftrightarrow E_{11}$	Add $E_7 \leftrightarrow E_9$ $E_8 \leftrightarrow E_{11}$

Delete/add $a \rightarrow b$ means that the path from a to b is deleted/added; delete/add $a \leftrightarrow b$ means that the covariance of a and b is deleted/added.

In McDonald's work, the covariance matrix was fitted by the model specified in Figure 1. Seven candidate models are constructed to fit the simulated data under each setting. The model displayed in Figure 1 is also in the candidate set, which we call α_4. Other models are constructed by deleting one or more parameters in α_4 (α_1, α_2, and α_3) or by adding one or more parameters to α_4 (α_5, α_6, and α_7). The detailed specifications of the other candidate models are given in Table 1. Note that the MDF and RMSEA values of these candidate models depend on which setting is considered. In each setting, the empirical probabilities of each candidate model selected by AIC, BIC, the RMSEA, and the RMSEA-2S are evaluated. Six levels of sample sizes are considered: 100, 200, 400, 800, 1600, and 6400. Under each sample size condition, the empirical probabilities are calculated based on 500 successful replications. All of the data sets are generated according to the method of Fleishman (1978) with the specified covariance structure under the corresponding setting. To see the behavior of AIC, BIC, and RMSEA under non-normality, the skewness and (Pearson's) kurtosis of each variable are set at 0 and 7, respectively.

In setting A, the population covariance matrix for simulating data is just the sample covariance matrix of McDonald (2010). The population MDF values, number of parameters, and RMSEA are presented in Table 2a. We can observe that α_7 has the smallest MDF value, but α_4 has the smallest RMSEA value. Hence, $\mathcal{A}_d = \mathcal{A}_d^* = \{\alpha_7\}$, and $\mathcal{A}_e = \mathcal{A}_e^* = \{\alpha_4\}$. Since all the population RMSEA values are larger than $c = .05$, we have $\mathcal{A}_c = \mathcal{A}_c^* = \emptyset$. Based on the derived theorems, both AIC and BIC are consistent for α_7; the RMSEA is consistent for α_4, and the RMSEA-2S selects nothing. The simulation results confirm the theoretical predictions (see Table 2b). Despite the fact that the difference between α_4 and α_7 is only slight in terms of both the population MDF and the RMSEA values, the considered model selection criteria could still differentiate them under the largest sample size. It is worth mentioning that both AIC and BIC prefer α_4 under small and moderate sample sizes, which indicates that the small sample performance of these selection criteria may differ from their large sample behavior.

Under Setting B, the true covariance matrix is the model-implied covariance obtained by fitting α_4 to the population covariance in Setting A. Hence, α_4 is of course a correct model, and so are α_5, α_6, and α_7, as shown in Table 3a. The corresponding optimal sets are $\mathcal{A}_d = \mathcal{A}_e = \{\alpha_4, \alpha_5, \alpha_6, \alpha_7\}$, $\mathcal{A}_d^* = \mathcal{A}_e^* = \{\alpha_4\}$, $\mathcal{A}_c = \{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$, and $\mathcal{A}_c^* = \{\alpha_2\}$. According to our theorems, we expect that asymptotically, BIC selects α_4; AIC and the RMSEA choose some model in $\mathcal{A}_d$ (or $\mathcal{A}_e$, since the two sets are the same), and the RMSEA-2S selects α_2. The empirical results support our prediction (see Table 3b). BIC selects α_4 with a near one probability under moderate and large sample sizes. AIC and the RMSEA choose α_4 with relatively high probability but could still choose α_5, α_6, and α_7, even when the sample size is quite large. Finally, the RMSEA-2S consistently chooses α_2 under both moderate and large sample sizes.

Under Setting C, the data generation process is quite similar to that of Setting B, but the population covariance matrix is slightly perturbed. Several error covariances are now set as 0.1, including the covariances of E_1 and E_4, E_2 and E_5, and E_3 and E_6. The population MDF and RMSEA values are presented in Table 4a. None of the candidate models are correct, and the

TABLE 2.
Model information and selection result of AIC, BIC, the RMSEA, and the RMSEA-2S in setting A.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
a. Model information							
$\mathcal{D}^*(\alpha)$	1.3113	1.1264	1.2682	1.0782	1.0761	1.0753	1.0717
$\lvert\alpha\rvert$	34	35	35	36	37	37	38
$\sqrt{\mathcal{D}^*(\alpha)/df(\alpha)}$	0.1359	0.1269	0.1346	0.1250	0.1258	0.1258	0.1265
b. Selection probability							
AIC							
$N = 100$	0.0080	0.1620	0.0000	0.3800	0.1660	0.2080	0.0760
$N = 200$	0.0000	0.0540	0.0000	0.3960	0.1920	0.2020	0.1560
$N = 400$	0.0000	0.0120	0.0000	0.3640	0.1700	0.2020	0.2520
$N = 800$	0.0000	0.0040	0.0000	0.2320	0.1760	0.1840	0.4040
$N = 1600$	0.0000	0.0000	0.0000	0.0680	0.1120	0.1220	0.6980
$N = 6400$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0060	0.9940
BIC							
$N = 100$	0.0240	0.4100	0.0120	0.4300	0.0460	0.0700	0.0080
$N = 200$	0.0000	0.2040	0.0000	0.6540	0.0480	0.0700	0.0240
$N = 400$	0.0000	0.0700	0.0000	0.7200	0.0640	0.1000	0.0460
$N = 800$	0.0000	0.0040	0.0000	0.6900	0.0840	0.1180	0.1040
$N = 1600$	0.0000	0.0000	0.0000	0.4860	0.1240	0.1920	0.1980
$N = 6400$	0.0000	0.0000	0.0000	0.0140	0.0200	0.0460	0.9200
RMSEA							
$N = 100$	0.0080	0.2460	0.0020	0.4240	0.1260	0.1620	0.0320
$N = 200$	0.0000	0.1440	0.0000	0.6320	0.0840	0.1020	0.0380
$N = 400$	0.0000	0.1080	0.0000	0.7540	0.0440	0.0700	0.0240
$N = 800$	0.0000	0.0320	0.0000	0.9060	0.0180	0.0360	0.0080
$N = 1600$	0.0000	0.0060	0.0000	0.9880	0.0000	0.0040	0.0020
$N = 6400$	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
RMSEA-2S							
$N = 100$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 200$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 400$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 800$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 1600$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 6400$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$\mathcal{D}^*(\alpha)$ is the population minimum function (MDF) value of model α; $|\alpha|$ is the number of parameters for model α; $df(\alpha)$ is the degrees of freedom of model α; the RMSEA-2S is the two-stage decision for the RMSEA.

optimal sets are $\mathcal{A}_d = \{\alpha_4, \alpha_5, \alpha_6, \alpha_7\}$, $\mathcal{A}_e = \mathcal{A}_d^* = \mathcal{A}_e^* = \{\alpha_4\}$, $\mathcal{A}_c = \{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$, and $\mathcal{A}_c^* = \{\alpha_2\}$. Our theorems posit that both BIC and the RMSEA will consistently select α_4; AIC will only choose a model in $\mathcal{A}_d$, and the RMSEA-2S will select α_2. The numerical results support our predictions for all the criteria (see Table 4b).

In Setting D, the population covariance is based on that of Setting C: the values of the covariances of E_1 and E_4, E_2 and E_5, and E_3 and E_6 are still 0.1, but the covariance of both E_7 and E_9 is set as 0.0107445. Table 5a shows that $\mathcal{A}_d = \{\alpha_5, \alpha_7\}$; $\mathcal{A}_d^* = \{\alpha_5\}$; $\mathcal{A}_e = \{\alpha_4, \alpha_5\}$; $\mathcal{A}_e^* = \{\alpha_4\}$; $\mathcal{A}_c = \{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$, and $\mathcal{A}_c^* = \{\alpha_2\}$. Our theoretical results imply that AIC is consistent for α_5 and α_7; BIC consistently chooses α_5; the RMSEA is consistent for α_4 and α_5, and the RMSEA-2S selects α_2. Our predictions are mostly supported except for

TABLE 3.
Model information and selection result of AIC, BIC, the RMSEA, and the RMSEA-2S in setting B.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
a. Model information							
$\mathcal{D}^*(\alpha)$	0.1327	0.0281	0.0982	0.0000	0.0000	0.0000	0.0000
$\lvert\alpha\rvert$	34	35	35	36	37	37	38
$\sqrt{\mathcal{D}^*(\alpha)/\mathrm{d}f(\alpha)}$	0.0611	0.0283	0.0530	0.0000	0.0000	0.0000	0.0000
b. Selection probability							
AIC							
$N = 100$	0.0000	0.1420	0.0040	0.5040	0.1220	0.1580	0.0700
$N = 200$	0.0000	0.0200	0.0000	0.6000	0.1640	0.1620	0.0540
$N = 400$	0.0000	0.0020	0.0000	0.5980	0.1500	0.1820	0.0680
$N = 800$	0.0000	0.0000	0.0000	0.5860	0.1680	0.1940	0.0520
$N = 1600$	0.0000	0.0000	0.0000	0.6100	0.1760	0.1680	0.0460
$N = 6400$	0.0000	0.0000	0.0000	0.5700	0.1580	0.2160	0.0560
BIC							
$N = 100$	0.0100	0.3860	0.0220	0.4940	0.0320	0.0520	0.0040
$N = 200$	0.0000	0.1440	0.0000	0.7800	0.0240	0.0480	0.0040
$N = 400$	0.0000	0.0280	0.0000	0.8880	0.0260	0.0580	0.0000
$N = 800$	0.0000	0.0000	0.0000	0.9580	0.0220	0.0200	0.0000
$N = 1600$	0.0000	0.0000	0.0000	0.9480	0.0300	0.0200	0.0020
$N = 6400$	0.0000	0.0000	0.0000	0.9700	0.0120	0.0180	0.0000
RMSEA							
$N = 100$	0.0080	0.1480	0.0020	0.3860	0.1660	0.1880	0.1020
$N = 200$	0.0000	0.0820	0.0000	0.4640	0.1780	0.1900	0.0860
$N = 400$	0.0000	0.0140	0.0000	0.4840	0.1640	0.2180	0.1200
$N = 800$	0.0000	0.0000	0.0000	0.5020	0.1880	0.2360	0.0740
$N = 1600$	0.0000	0.0000	0.0000	0.5300	0.2080	0.1620	0.1000
$N = 6400$	0.0000	0.0000	0.0000	0.4600	0.1900	0.2400	0.1100
RMSEA-2S							
$N = 100$	0.1180	0.3020	0.0060	0.1060	0.0140	0.0060	0.0020
$N = 200$	0.0720	0.6160	0.0020	0.1580	0.0040	0.0140	0.0020
$N = 400$	0.0300	0.9340	0.0000	0.0360	0.0000	0.0000	0.0000
$N = 800$	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 1600$	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$N = 6400$	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$\mathcal{D}^*(\alpha)$ is the population minimum function (MDF) value of model α; $|\alpha|$ is the number of parameters for model α; $df(\alpha)$ is the degrees of freedom of model α; the RMSEA-2S is the two-stage decision for the RMSEA.

the performance of BIC (see Table 5b). BIC tends to select α_4 when the sample sizes are between 200 and 1600. Even with the largest sample size, BIC still selects α_4 with a probability of 0.54. We speculate that the sample size of 6400 is not sufficient for demonstrating the consistency of BIC for α_5. Hence, an additional simulation with a sample size of 25600 is conducted. The result shows that the probability of selecting α_5 by BIC is now 0.966. The consistency of BIC is still observed, although a super large sample size is required.

5. Discussion

In this study, the asymptotic behaviors of AIC, BIC, and RMSEA under nested and non-nested model selection are derived. An advantage of our results is that it does not depend on the

TABLE 4.
Model information and selection result of AIC, BIC, the RMSEA, and the RMSEA-2S in setting C.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
a. Model information							
$\mathcal{D}^*(\alpha)$	0.1679	0.0682	0.1457	0.0440	0.0440	0.0440	0.0440
$\lvert\alpha\rvert$	34	35	35	36	37	37	38
$\sqrt{\mathcal{D}^*(\alpha)/df(\alpha)}$	0.0688	0.0441	0.0645	0.0357	0.0360	0.0360	0.0362
b. Selection probability							
AIC							
$N = 100$	0.0020	0.1900	0.0060	0.4820	0.1100	0.1720	0.0380
$N = 200$	0.0000	0.0500	0.0000	0.5480	0.1520	0.1940	0.0560
$N = 400$	0.0000	0.0020	0.0000	0.6140	0.1740	0.1640	0.0460
$N = 800$	0.0000	0.0000	0.0000	0.6460	0.1440	0.1600	0.0500
$N = 1600$	0.0000	0.0000	0.0000	0.6020	0.1840	0.1600	0.0540
$N = 6400$	0.0000	0.0000	0.0000	0.5780	0.1580	0.2080	0.0560
BIC							
$N = 100$	0.0120	0.4340	0.0140	0.4620	0.0200	0.0540	0.0040
$N = 200$	0.0000	0.2220	0.0000	0.6860	0.0360	0.0540	0.0020
$N = 400$	0.0000	0.0360	0.0000	0.8940	0.0280	0.0420	0.0000
$N = 800$	0.0000	0.0000	0.0000	0.9640	0.0140	0.0200	0.0020
$N = 1600$	0.0000	0.0000	0.0000	0.9580	0.0200	0.0220	0.0000
$N = 6400$	0.0000	0.0000	0.0000	0.9820	0.0040	0.0140	0.0000
RMSEA							
$N = 100$	0.0100	0.1260	0.0020	0.4140	0.1600	0.2240	0.0640
$N = 200$	0.0000	0.0320	0.0000	0.4700	0.1820	0.2320	0.0840
$N = 400$	0.0000	0.0020	0.0000	0.5460	0.1880	0.1940	0.0700
$N = 800$	0.0000	0.0000	0.0000	0.6860	0.1340	0.1400	0.0400
$N = 1600$	0.0000	0.0000	0.0000	0.7420	0.1280	0.1180	0.0120
$N = 6400$	0.0000	0.0000	0.0000	0.9840	0.0040	0.0120	0.0000
RMSEA-2S							
$N = 100$	0.6840	0.1880	0.0080	0.0900	0.0080	0.0100	0.0000
$N = 200$	0.4460	0.3500	0.0000	0.1760	0.0080	0.0120	0.0000
$N = 400$	0.1580	0.5380	0.0000	0.2860	0.0040	0.0020	0.0000
$N = 800$	0.0220	0.7180	0.0000	0.2600	0.0000	0.0000	0.0000
$N = 1600$	0.0000	0.9020	0.0000	0.0980	0.0000	0.0000	0.0000
$N = 6400$	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$\mathcal{D}^*(\alpha)$ is the population minimum function (MDF) value of model α; $|\alpha|$ is the number of parameters for model α; $df(\alpha)$ is the degrees of freedom of model α; the RMSEA-2S is the two-stage decision for the RMSEA.

distributional form of the data and the existence of a true model in the candidate set. Therefore, the derived results can be applied to most SEM applications. From the point of view of SEM users, one may ask which criterion should be used in practice. We believe that the answer depends on the purpose of model selection. If the researcher hopes to select a model with smallest population MDF value, AIC and BIC are better for this purpose. In particular, the derived theorem shows that BIC can consistently select the most parsimonious one from all of the models that attains the smallest MDF value given the assumption that the implied covariance matrices of these models are identical. On the other hand, when the researcher hopes to choose a model with the smallest population RMSEA, the RMSEA is better. The RMSEA-2S is mostly appropriate if researchers

TABLE 5.
Model information and selection result of AIC, BIC, the RMSEA, and the RMSEA-2S in setting D.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
a. Model information							
$\mathcal{D}^*(\alpha)$	0.1689	0.0688	0.1481	0.0447	0.0440	0.0446	0.0440
$\lvert\alpha\rvert$	34	35	35	36	37	37	38
$\sqrt{\mathcal{D}^*(\alpha)/df(\alpha)}$	0.0690	0.0443	0.0650	0.0360	0.0360	0.0362	0.0362
b. Selection probability							
AIC							
$N = 100$	0.0000	0.1660	0.0080	0.4660	0.1540	0.1680	0.0380
$N = 200$	0.0000	0.0420	0.0000	0.5320	0.1760	0.1800	0.0700
$N = 400$	0.0000	0.0000	0.0000	0.5220	0.2300	0.1920	0.0560
$N = 800$	0.0000	0.0000	0.0000	0.5020	0.2680	0.1480	0.0820
$N = 1600$	0.0000	0.0000	0.0000	0.3840	0.3800	0.1180	0.1180
$N = 6400$	0.0000	0.0000	0.0000	0.0860	0.6420	0.0320	0.2400
BIC							
$N = 100$	0.0080	0.4420	0.0140	0.4360	0.0540	0.0420	0.0040
$N = 200$	0.0000	0.2120	0.0000	0.6820	0.0360	0.0620	0.0080
$N = 400$	0.0000	0.0400	0.0000	0.8420	0.0780	0.0380	0.0020
$N = 800$	0.0000	0.0000	0.0000	0.9060	0.0660	0.0280	0.0000
$N = 1600$	0.0000	0.0000	0.0000	0.8480	0.1280	0.0220	0.0020
$N = 6400$	0.0000	0.0000	0.0000	0.5620	0.4140	0.0160	0.0080
RMSEA							
$N = 100$	0.0020	0.1360	0.0100	0.3820	0.1800	0.2140	0.0760
$N = 200$	0.0000	0.0380	0.0000	0.4560	0.2160	0.1900	0.1000
$N = 400$	0.0000	0.0020	0.0000	0.4900	0.2420	0.1860	0.0800
$N = 800$	0.0000	0.0000	0.0000	0.5360	0.2540	0.1440	0.0660
$N = 1600$	0.0000	0.0000	0.0000	0.5580	0.3100	0.0880	0.0440
$N = 6400$	0.0000	0.0000	0.0000	0.5940	0.3880	0.0140	0.0040
RMSEA-2S							
$N = 100$	0.0560	0.2200	0.0060	0.0680	0.0240	0.0080	0.0040
$N = 200$	0.0180	0.3440	0.0000	0.1900	0.0120	0.0140	0.0040
$N = 400$	0.0000	0.5360	0.0000	0.3100	0.0160	0.0040	0.0000
$N = 800$	0.0000	0.7160	0.0000	0.2560	0.0060	0.0000	0.0000
$N = 1600$	0.0000	0.9100	0.0000	0.0880	0.0000	0.0000	0.0000
$N = 6400$	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$\mathcal{D}^*(\alpha)$ is the population minimum function (MDF) value of model α; $|\alpha|$ is the number of parameters for model α; $df(\alpha)$ is the degrees of freedom of model α; the RMSEA-2S is the two-stage decision for the RMSEA.

hope to find the most parsimonious one from a set of models with reasonable fit in terms of population RMSEA.

The consistency of BIC for choosing a quasi-true model (i.e., a model in $\mathcal{A}_d^*$) may be strange to people who have heard "BIC is not consistent if the true model is not in the candidate set, or the true model is nonparametric" (e.g., Vrieze, 2010). Actually, our results are not contrary to the existing results. One reason for this is that our definition of consistency is different. Theorem 2 shows that BIC is consistent for a quasi-true model as defined by the MDF value and number of parameters, but not the true model or the optimal model minimizing the sample-dependent loss. Another reason is that SEM is a pure parametric method. If we assume that the population covariance matrix is just an unknown but fixed quantity, it can always be perfectly explained by some models with $P^* = P\left(P+1\right)/2$ parameters.

The established results can be applied to any discrepancy function that can be quadratically approximated. Hence, even when an estimation method other than ML is used, researchers still have AIC/BIC/RMSEA-type criteria that can consistently select a model in some optimal set. Note that the chosen discrepancy not only determines the estimates but also the corresponding optimal sets. In general, selection criteria based on discrepancy A cannot select an optimal model defined by discrepancy B unless some candidate model is correct.

The current study also presents numerical illustrations for demonstrating the finite sample and asymptotic behaviors of AIC, BIC, the RMSEA, and the RMSEA-2S. Because SEM users always work with imperfect models (MacCallum, 2003), we adopt an empirical covariance matrix from McDonald (2010) as the target population covariance to enhance the ecological validity of the demonstration. In general, the illustrations support our theoretical results, despite the fact that in some settings extremely large sample sizes are required to achieve the limiting behaviors, especially for the case of BIC. However, the numerical illustrations utilize "population covariance matrices" with "sampling errors." The true model underlying the target covariance matrix is actually unknown. If someone believes that psychological data should be generated according to some well-defined true model, our approach may not be appropriate, and the simulation results should be cautiously interpreted. In future, it is worth exploring the empirical performances of these criteria under a well-controlled manipulation of underlying true models.

The main limitation of the current study is that our analysis relies on the asymptotic theory. Under finite sample sizes, however, the empirical performances of model selection criteria can be quite different from their limiting behaviors, as shown in our numerical illustrations. Vrieze (2012) also showed that BIC cannot select the true model under small parameter values even when the sample size is quite large. Another issue related to finite sample sizes is model selection uncertainty. Ignoring the fact that model selection uncertainty can lead to invalid inferences (Preacher & Merkle, 2012). Further comprehensive simulations are required to see the entire picture of the behaviors of these selection criteria. Another limitation is that we only consider model selection problems with complete data. Missing data are easily encountered in practice. Since sample size is not well defined in the presence of missing data, the asymptotic behaviors of the selection criteria could not be directly analyzed. Further research should study the issue of model selection with missing data in SEM. We believe that the criteria proposed by Ibrahim, Zhu, and Tang (2008) is a promising approach for such problems.

Appendix

The following two lemmas are helpful for proving the four main theorems.

Lemma 1. *Let $\mathcal{G}_N$ denote a random function of α and $\mathcal{B} = \left\{\max_{\alpha_1 \in \mathcal{A}_1} \mathcal{G}_N\left(\alpha_1\right) < \min_{\alpha_2 \in \mathcal{A}_2} \mathcal{G}_N\left(\alpha_2\right)\right\}$. If the cardinality of $\mathcal{A}_1$ and $\mathcal{A}_2$ are both finite, and $\mathbb{P}\left(\mathcal{G}_N\left(\alpha_1\right) > \mathcal{G}_N\left(\alpha_2\right)\right) \to 0$ for each $\alpha_1 \in \mathcal{A}_1$ and $\alpha_2 \in \mathcal{A}_2$, then*

$$\mathbb{P}\left(\mathcal{B}\right) \to 1.$$

Proof of Lemma 1. It suffices to show that the probability of $\mathcal{B}^c$, the complement of $\mathcal{B}$, converges to zero. By the fact $\mathcal{B}^c \subset \bigcup_{\alpha_1 \in \mathcal{A}_1, \alpha_2 \in \mathcal{A}_2} \left\{\mathcal{G}_N\left(\alpha_1\right) > \mathcal{G}_N\left(\alpha_2\right)\right\}$, Boole's inequality implies that

$$\mathbb{P}\left(\mathcal{B}^c\right) \leq \sum_{\alpha_1 \in \mathcal{A}_1, \alpha_2 \in \mathcal{A}_2} \mathbb{P}\left(\mathcal{G}_N\left(\alpha_1\right) > \mathcal{G}_N\left(\alpha_2\right)\right).$$

Since both $\mathcal{A}_1$ and $\mathcal{A}_2$ are finite, and each $\mathbb{P}\left(\mathcal{G}_N(\alpha_1) > \mathcal{G}_N(\alpha_2)\right) \to 0$, the right-hand side converges to zero as $N \to +\infty$.

Lemma 1 implies that under finite $\mathcal{A}$, if we can show that $\mathbb{P}\left(\mathcal{C}(\alpha_1, \mathcal{D}, s) > \mathcal{C}(\alpha_2, \mathcal{D}, s)\right) \to 0$ for each $\alpha_1 \in \mathcal{A}_1$ and $\alpha_2 \in \mathcal{A}_2$, then $\hat{\alpha}_N \in \mathcal{A}_1$. □

Lemma 2. *We define $\mathcal{F}^*(\alpha) = \frac{\partial^2 \mathcal{D}(\sigma^*(\alpha), \sigma^0)}{\partial\theta_\alpha \partial\theta_\alpha^T}$ and $\mathcal{J}^*(\alpha) = \frac{\partial \mathcal{D}(\sigma^*(\alpha), \sigma^0)}{\partial\theta_\alpha \partial\sigma^T}$. Let α_1 and α_2 denote two indexes of models. Consider two test statistics*

$$T_N(\alpha_1, \alpha_2, w_1, w_2) = N\left(w_1 \mathcal{D}\left(\hat{\sigma}(\alpha_1), s\right) - w_2 \mathcal{D}\left(\hat{\sigma}(\alpha_2), s\right)\right),$$

and

$$Z_N(\alpha_1, \alpha_2, w_1, w_2) = \sqrt{N}\left(w_1 \mathcal{D}\left(\hat{\sigma}(\alpha_1), s\right) - w_2 \mathcal{D}\left(\hat{\sigma}(\alpha_2), s\right)\right),$$

where w_1 and w_2 are two nonnegative weights.

(1) *If $w_1 \mathcal{D}\left(\sigma^*(\alpha_1), \sigma^0\right) = w_2 \mathcal{D}\left(\sigma^*(\alpha_2), \sigma^0\right)$, and $\sigma^*(\alpha_1) = \sigma^*(\alpha_2)$, but $|\alpha_1| < |\alpha_2|$,*

$$T_N(\alpha_1, \alpha_2, w_1, w_2) \longrightarrow_L T(\alpha_1, \alpha_2, w_1, w_2) = \sum_k \lambda_k \chi_k^2,$$

where χ_k^2's are independent chi-square random variables, and λ_k is the k^{th} eigenvalue of $\mathcal{W}^*(\alpha_1, \alpha_2, w_1, w_2)\,\mathcal{V}^*(\alpha_1, \alpha_2)$ with

$$\mathcal{W}^*(\alpha_1, \alpha_2, w_1, w_2) = \frac{1}{2}\begin{pmatrix} w_1 \mathcal{F}^*(\alpha_1) & 0 \\ 0 & -w_2 \mathcal{F}^*(\alpha_2) \end{pmatrix}$$

and

$$\begin{aligned} &\mathcal{V}^*(\alpha_1, \alpha_2) \\ &= \begin{pmatrix} \mathcal{F}^*(\alpha_1)^{-1} \mathcal{J}^*(\alpha_1)\,\Gamma \mathcal{J}^*(\alpha_1)^T \mathcal{F}^*(\alpha_1)^{-1} & \\ \mathcal{F}^*(\alpha_2)^{-1} \mathcal{J}^*(\alpha_2)\,\Gamma \mathcal{J}^*(\alpha_1)^T \mathcal{F}^*(\alpha_1)^{-1} & \mathcal{F}^*(\alpha_2)^{-1} \mathcal{J}^*(\alpha_2)\,\Gamma \mathcal{J}^*(\alpha_2)^T \mathcal{F}^*(\alpha_2)^{-1} \end{pmatrix}. \end{aligned}$$

In particular, if $w_1 = w_2 = 1$, then $T_N(\alpha_1, \alpha_2) \equiv T_N(\alpha_1, \alpha_2, 1, 1) \longrightarrow_L T(\alpha_1, \alpha_2)$.

(2) *If $w_1 \mathcal{D}\left(\sigma^*(\alpha_1), \sigma^0\right) = w_2 \mathcal{D}\left(\sigma^*(\alpha_2), \sigma^0\right)$, but $\sigma^*(\alpha_1) \neq \sigma^*(\alpha_2)$, then*

$$Z_N(\alpha_1, \alpha_2, w_1, w_2) \longrightarrow_L Z(\alpha_1, \alpha_2, w_1, w_2),$$

where $Z(\alpha_1, \alpha_2, w_1, w_2) \sim N\left(0, \nu(\alpha_1, \alpha_2, w_1, w_2)^T \Gamma \nu(\alpha_1, \alpha_2, w_1, w_2)\right)$, with Γ being the limiting covariance of $\sqrt{N}\left(s - \sigma^0\right)$, and

$$\nu(\alpha_1, \alpha_2, w_1, w_2) = w_1 \frac{\partial \mathcal{D}\left(\sigma^*(\alpha_1), \sigma^0\right)}{\partial\sigma} - w_2 \frac{\partial \mathcal{D}\left(\sigma^*(\alpha_2), \sigma^0\right)}{\partial\sigma}.$$

In particular, if $w_1 = w_2 = 1$, then $Z_N(\alpha_1, \alpha_2) \equiv Z_N(\alpha_1, \alpha_2, 1, 1) \longrightarrow_L Z(\alpha_1, \alpha_2)$.

Lemma 2 can be seen as a variant of Theorem 3.3 from Vuong (1989) under the SEM settings with general discrepancy function $\mathcal{D}$. The proof of part (1) relies on the consistency and the asymptotic distribution of an MDF estimator under misspecified SEM models (see Satorra, 1989; Shapiro, 1983, 1984, 2007). Similar results can be also found in Satorra and Bentler (2001). Part (2) can be justified by the Delta method if we treat the discrepancy function as a function of a sample covariance vector (see Shapiro, 2009 for more general results). The complete proof of Lemma 2 can be found in the online supplemental material.

Because the consistency of the MDF estimator is crucial for deriving our results, the technical details of Theorem 1 in Shapiro (1984) are briefly discussed here. The consistency of an MDF estimator depends on the following: (a) $\mathcal{D}(\sigma_\alpha(\theta_\alpha), \sigma)$ is a continuous function in both θ_α and σ; (b) Θ_α is compact; (c) θ_α is conditionally identified at $\theta_\alpha^* \in \Theta_\alpha$, given $\sigma = \sigma^0$; (d) s is a consistent estimator for σ. Obviously, (a) is implied by our conditions C and D. (b) is satisfied by the part (2) of Condition E. Part (1) of Condition E implies (c) to be true. Finally, (d) can be obtained by using Condition A. Shapiro (1984) also observed that in practice the compactness of Θ_α does not hold. Hence, Shapiro proposed the condition of inf-boundedness: There exists a $\delta > \mathcal{D}\left(\sigma_\alpha\left(\theta_\alpha^*\right), \sigma^0\right)$ and a compact subset $\Theta_\alpha^* \subset \Theta_\alpha$ such that $\{\theta_\alpha | \mathcal{D}(\sigma_\alpha(\theta_\alpha), \sigma) < \delta\} \subset \Theta_\alpha^*$ whenever σ is in the neighborhood of σ^0. Under this condition, the minimization actually takes place on Θ_α^* for all σ near σ^0. Although it may not be easy to justify the inf-boundedness condition for all types of SEM models, finding a counterexample of practical interest is also difficult.

Proof of Theorem 1

(1) If $\mathcal{A}_d = \mathcal{A}$, part (1) holds trivially. For $\mathcal{A} \backslash \mathcal{A}_d \neq \emptyset$, by Lemma 1, we only need to show

$$\mathbb{P}\left(IC_{k_N}(\alpha_d) > IC_{k_N}(\alpha)\right) \to 0,$$

for each $\alpha_d \in \mathcal{A}_d$ and $\alpha \in \mathcal{A} \backslash \mathcal{A}_d$. Since $IC_{k_N}(\alpha_d) \longrightarrow_P \mathcal{D}^*(\alpha_d)$ and $IC_{k_N}(\alpha) \longrightarrow_P \mathcal{D}^*(\alpha) > \mathcal{D}^*(\alpha_d)$ under $k_N = O_{\mathbb{P}}\left(N^{-1}\right)$, given $\epsilon > 0$ we can find $N(\epsilon)$ such that $\mathbb{P}\left(IC_{k_N}(\alpha_d) > \frac{\mathcal{D}^*(\alpha)+\mathcal{D}^*(\alpha_d)}{2}\right) < \frac{\epsilon}{2}$ and $\mathbb{P}\left(IC_{k_N}(\alpha) < \frac{\mathcal{D}^*(\alpha)+\mathcal{D}^*(\alpha_d)}{2}\right) < \frac{\epsilon}{2}$ whenever $N > N(\epsilon)$. Hence, we have $\mathbb{P}\left(IC_{k_N}(\alpha_d) > IC_{k_N}(\alpha)\right) < \epsilon$ if $N > N(\epsilon)$.

(2) Let α denote any element in $\mathcal{A}_d \backslash \alpha_d^*$. Since the event $\left\{IC_{k_N}\left(\alpha_d^*\right) - IC_{k_N}(\alpha) > 0\right\}$ is contained in $\left\{\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right\}$, we have $\mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) \geq \mathbb{P}\left(IC_{k_N}\left(\alpha_d^*\right) - IC_{k_N}(\alpha) > 0\right)$.

Case A: $\sigma^*\left(\alpha_d^*\right) = \sigma^*(\alpha)$. The assumption implies that $N\left(IC_{k_N}(\alpha_d) - IC_{k_N}(\alpha)\right) = T_N\left(\alpha_d^*, \alpha\right) + Nk_N\left(\left|\alpha_d^*\right| - |\alpha|\right)$. Since $\lim_{N\to\infty} \mathbb{P}(Nk_N \leq M) = 1$ for some $M < +\infty$ by the fact $k_N = O_{\mathbb{P}}\left(N^{-1}\right)$, we have

$$\mathbb{P}\left(N\left(IC_{k_N}(\alpha_d) - IC_{k_N}(\alpha)\right) > 0\right) \to \mathbb{P}\left(T\left(\alpha_d^*, \alpha\right) > M\left(\left|\alpha_d^*\right| - |\alpha|\right)\right) > 0,$$

and conclude $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) \geq \max_{\alpha \in \mathcal{A}_d \backslash \alpha_d^*} \mathbb{P}\left(T\left(\alpha_d^*, \alpha\right) > M\left(\left|\alpha_d^*\right| - |\alpha|\right)\right) > 0$.

Case B: $\sigma^*\left(\alpha_d^*\right) \neq \sigma^*(\alpha)$. Since $\sqrt{N}\left(IC_{k_N}(\alpha_d) - IC_{k_N}(\alpha)\right) = Z_N\left(\alpha_d^*, \alpha\right) + \sqrt{N}k_N\left(\left|\alpha_d^*\right| - |\alpha|\right)$, we have

$$\mathbb{P}\left(\sqrt{N}\left(IC_{k_N}(\alpha_d) - IC_{k_N}(\alpha)\right) > 0\right) \to \mathbb{P}\left(Z\left(\alpha_d^*, \alpha\right) > 0\right) > 0.$$

Therefore, $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) \geq \max_{\alpha \in \mathcal{A}_d \backslash \alpha_d^*} \mathbb{P}\left(Z\left(\alpha_d^*, \alpha\right) > 0\right) > 0.$ □

Proof of Theorem 2.

(1) Let $\alpha_d \in \mathcal{A}_d$ and $\alpha \in \mathcal{A} \backslash \mathcal{A}_d$.

$$\begin{aligned} \mathbb{P}\left(IC_{k_N}(\alpha_d) - IC_{k_N}(\alpha) > 0\right) &= \mathbb{P}\left(\hat{\mathcal{D}}(\alpha_d) - \hat{\mathcal{D}}(\alpha) + k_N(|\alpha_d| - |\alpha|) > 0\right) \\ &\to \mathbb{P}\left(\mathcal{D}^*(\alpha_d) - \mathcal{D}^*(\alpha) > 0\right) = 0. \end{aligned}$$

(2) For each $\alpha \in \mathcal{A}_d \backslash \mathcal{A}_d^*$, we have

$$\begin{aligned} \mathbb{P}\left(IC_{k_N}\left(\alpha_d^*\right) - IC_{k_N}(\alpha) > 0\right) &= \mathbb{P}\left(N\left(IC_{k_N}\left(\alpha_d^*\right) - IC_{k_N}(\alpha)\right) > 0\right) \\ &= \mathbb{P}\left(T_N\left(\alpha_d^*, \alpha\right) > Nk_N\left(|\alpha| - \left|\alpha_d^*\right|\right)\right) \\ &\longrightarrow \mathbb{P}\left(T\left(\alpha_d^*, \alpha\right) > +\infty\right) = 0. \end{aligned}$$

By lemma 1, we conclude $\mathbb{P}\left(IC_{k_N}\left(\alpha_d^*\right) > \min_{\alpha \in \mathcal{A}_d \backslash \alpha_d^*} IC_{k_N}(\alpha)\right) \longrightarrow 0$ and $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N = \alpha_d^*\right) = 1$.

(3) Choose $\alpha \in \mathcal{A}_d \backslash \alpha_d^*$, then

$$\begin{aligned} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) &\geq \mathbb{P}\left(\sqrt{N}\left(IC_{k_N}\left(\alpha_d^*\right) - IC_{k_N}(\alpha)\right) > 0\right) \\ &= \mathbb{P}\left(Z\left(\alpha_d^*, \alpha\right) > \sqrt{N}k_N\left(|\alpha| - \left|\alpha_d^*\right|\right) + o_{\mathbb{P}}(1)\right) \longrightarrow \mathbb{P}\left(Z\left(\alpha_d^*, \alpha\right)\right. \\ &\left.> M\left(|\alpha| - \left|\alpha_d^*\right|\right)\right) \end{aligned}$$

Therefore, $\lim_{N\to\infty} \mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_d \backslash \alpha_d^*\right) \geq \max_{\alpha \in \mathcal{A}_d \backslash \alpha_d^*} \mathbb{P}\left(Z\left(\alpha_d^*, \alpha\right) > M\left(|\alpha| - \left|\alpha_d^*\right|\right)\right) > 0$.

□

Proof of Theorem 3.

(1) Let $\alpha_e \in \mathcal{A}_e$ and $\alpha \in \mathcal{A} \backslash \mathcal{A}_e$. Because $\frac{\hat{\mathcal{D}}(\alpha_e)}{df(\alpha_e)} - \frac{1}{N} \longrightarrow_P \frac{\mathcal{D}^*(\alpha_e)}{df(\alpha_e)}$ and $\frac{\hat{\mathcal{D}}(\alpha)}{df(\alpha)} - \frac{1}{N} \longrightarrow_P \frac{\mathcal{D}^*(\alpha)}{df(\alpha)} > \frac{\mathcal{D}^*(\alpha_e)}{df(\alpha_e)}$, we have

$$\mathbb{P}\left(RMSEA_N(\alpha_e) - RMSEA_N(\alpha) > 0\right) = \mathbb{P}\left(\frac{\mathcal{D}^*(\alpha_e)}{df(\alpha_e)} > \frac{\mathcal{D}^*(\alpha)}{df(\alpha)} + o_{\mathbb{P}}(1)\right) \longrightarrow 0.$$

(2) Let $\alpha \in \mathcal{A}_e \backslash \mathcal{A}_e^*$. By the definition of α_e^* and $\mathcal{A}_e \backslash \alpha_e^*$, we know that $\frac{\mathcal{D}^*(\alpha_e^*)}{df(\alpha_e^*)} = \frac{\mathcal{D}^*(\alpha)}{df(\alpha)}$ and hence $df(\alpha)\mathcal{D}^*\left(\alpha_e^*\right) = df\left(\alpha_e^*\right)\mathcal{D}^*(\alpha)$. Since the event $\left\{RMSEA_N\left(\alpha_e^*\right) - RMSEA_N(\alpha) > 0\right\}$ is contained in $\left\{\hat{\alpha}_N \in \mathcal{A}_e \backslash \alpha_e^*\right\}$, we have $\mathbb{P}\left(\hat{\alpha}_N \in \mathcal{A}_e \backslash \alpha_e^*\right) \geq \mathbb{P}\left(RMSEA_N\left(\alpha_e^*\right) - RMSEA_N(\alpha) > 0\right)$.

Case A. $df(\alpha)\mathcal{D}^*\left(\alpha_e^*\right) = df\left(\alpha_e^*\right)\mathcal{D}^*(\alpha) = 0$. Since the event $\left\{\left(\frac{\hat{\mathcal{D}}(\alpha_e^*)}{df(\alpha_e^*)} - \frac{1}{N}\right) - \frac{\hat{\mathcal{D}}(\alpha)}{df(\alpha)} > 0\right\}$ is contained in $\left\{\max\left\{\frac{\hat{\mathcal{D}}(\alpha_e^*)}{df(\alpha_e^*)} - \frac{1}{N}, 0\right\} - \max\left\{\frac{\hat{\mathcal{D}}(\alpha)}{df(\alpha)} - \frac{1}{N}, 0\right\} > 0\right\} = \left\{RMSEA_N\left(\alpha_e^*\right) - RMSEA_N(\alpha) > 0\right\}$, we have

$$\begin{aligned}\mathbb{P}\left(RMSEA_N\left(\alpha_e^*\right)-RMSEA_N\left(\alpha\right)>0\right) &\geq \mathbb{P}\left(\frac{\hat{\mathcal{D}}\left(\alpha_e^*\right)}{df\left(\alpha_e^*\right)}-\frac{\hat{\mathcal{D}}\left(\alpha\right)}{df\left(\alpha\right)}>\frac{1}{N}\right)\\ &=\mathbb{P}\left(T_N\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)\right.\\ &\left.> df\left(\alpha\right)df\left(\alpha_e^*\right)\right)\\ &\quad\rightarrow \mathbb{P}\left(T\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)\right.\\ &\left.> df\left(\alpha\right)df\left(\alpha_e^*\right)\right)>0\end{aligned}$$

Hence, $\lim_{N\rightarrow\infty}\mathbb{P}\left(\hat{\alpha}_N\in\mathcal{A}_e\backslash\alpha_e^*\right)\geq\max_{\alpha\in\mathcal{A}_e\backslash\alpha_e^*}\mathbb{P}\left(T\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)>0\right)>0$.
Case B. $df\left(\alpha\right)\mathcal{D}^*\left(\alpha_e^*\right)=df\left(\alpha_e^*\right)\mathcal{D}^*\left(\alpha\right)$ but $\sigma^*\left(\alpha_e^*\right)\neq\sigma^*\left(\alpha\right)$. Through similar technique in case A, we have

$$\begin{aligned}&\mathbb{P}\left(RMSEA_N\left(\alpha_e^*\right)-RMSEA_N\left(\alpha\right)>0\right)\geq\mathbb{P}\left(\frac{\hat{\mathcal{D}}\left(\alpha_e^*\right)}{df\left(\alpha_e^*\right)}-\frac{\hat{\mathcal{D}}\left(\alpha\right)}{df\left(\alpha\right)}>\frac{1}{N}\right)\\ &=\mathbb{P}\left(Z_N\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)>\frac{df\left(\alpha\right)df\left(\alpha_e^*\right)}{\sqrt{N}}\right)\\ &\rightarrow\mathbb{P}\left(Z\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)>0\right)>0\end{aligned}$$

We conclude that $\lim_{N\rightarrow\infty}\mathbb{P}\left(\hat{\alpha}_N\in\mathcal{A}_e\backslash\alpha_e^*\right)\geq\max_{\alpha\in\mathcal{A}_e\backslash\alpha_e^*}\mathbb{P}\left(Z\left(\alpha_e^*,\alpha,df\left(\alpha\right),df\left(\alpha_e^*\right)\right)>0\right)$ >0.

□

Proof of Theorem 4.

By the fact $\frac{\hat{\mathcal{D}}(\alpha)}{df(\alpha)}-\frac{1}{N}\longrightarrow_P\frac{\mathcal{D}^*(\alpha)}{df(\alpha)}$ for each $\alpha\in\mathcal{A}$ and $\frac{\mathcal{D}^*(\alpha_c)}{df(\alpha_c)}<c$ for all $\alpha_c\in\mathcal{A}_c$, we have

$$\mathbb{P}\left(\bigcup\nolimits_{\alpha_c\in\mathcal{A}_c}\left\{RMSEA_N\left(\alpha_c\right)>c\right\}\right)\leq\sum_{\alpha_c\in\mathcal{A}_c}\mathbb{P}\left(RMSEA_N\left(\alpha_c\right)>c\right)\longrightarrow 0$$

Hence, in the first stage, we can correctly identify all the models in $\mathcal{A}_c$ under large N. Since the second stage is just to compare $|\alpha_c|$ of each model in $\mathcal{A}_c$, a non-random quantity, we conclude that $\lim_{N\rightarrow\infty}\mathbb{P}\left(\hat{\alpha}_N\in\mathcal{A}_c^*\right)=1$. □

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, *45*, 289–308.

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*, *21*, 1–19.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

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

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–62). Newbury Park, CA: Sage.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria (Tech. Rep. No. 12–119). University Park, PA: The Pennsylvania State University, The Methodology Center.

Feist, G. J., Bodner, T. E., Jacobs, J. F., Miles, M., & Tan, V. (1995). Integrating top-down and bottom-up structural models of subjective well-being: A longitudinal investigation. *Journal of Personality and Social Psychology*, *68*, 138–150.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190–195.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, *16*, 342–355.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. Communication in Statistics. *Part B: Simulation and Computation*, *26*, 1477–1516.

Homburg, C. (1991). Cross-validation and information criteria in causal modeling. *Journal of Marketing Research*, *28*, 137–144.

Ibrahim, J. G., Zhu, H.-T., & Tang, N.-S. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, *103*, 1648–1658.

Jackson, D. L., Gillaspy, J. A, Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6–23.

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

Kaplan, D. (2009). *Structural Equation Modeling: Foundations and Extensions* (2nd ed.). Newbury Park, CA: SAGE Publications.

Keyes, C. L. M., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of Personality and Social Psychology*, *82*, 1007–1022.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Li, L. & Bentler, P. M. (2006). Robust statistical tests for evaluating the hypothesis of close fit of misspecified mean and covariance structural models. UCLA Statistics Preprint #494.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–224.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*, 675–686.

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

Pitt, M. A., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259.

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

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*, 28–56.

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

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, *24*, 148–169.

Shao, J. (1997). An asymptotic theory for model selection. *Statistics Sinica*, *7*, 221–264.

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures (a unified approach). *South African Statistical Journal*, *17*, 33–81.

Shapiro, A. (1984). A note on the consistency of estimators in the analysis of moment structures. *British Journal of Mathematical and Statistical Psychology*, *1984*, 84–88.

Shapiro, A. (2007). Statistical inference of moment structures. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 229–260). Amsterdam: Elsevier.

Shapiro, A. (2009). Asymptotic normality of test statistics under alternative hypotheses. *Journal of Multivariate Analysis*, *100*, 936–945.

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Iowa City, IA: Paper presented at the annual Spring Meeting of the Psychometric Society.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society, Series B*, *36*, 111–147.

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

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Manuscript Received: 12 SEP 2014
Final Version Received: 24 MAR 2017
Published Online Date: 26 APR 2017