Information and Other Criteria in Structural Equation Model Selection

Johan H.L. Oud
University of Nijmegen, The Netherlands

Dominique M.A. Haughton
Bentley College

Robert A.R.G. Jansen
University of Nijmegen, The Netherlands

Abstract

This article presents the results of a simulation study evaluating information criteria and other well-known criteria for model selection in structural equation modeling. In the presence of overfitting, underfitting, and correctly specified analytic models and using sample sizes of $n = 100, 400, 1000, 6000$, the performance of the criteria is assessed by the frequency each of the analytic models is selected as best by the criterion. We find that the information criteria perform better than other criteria overall, but that the cross-validation index remains an attractive option. Within the class of information criteria, the Akaike information criterion is found to show some overfitting tendency. Contrary to suggestions in the literature but in accordance with theoretical results, this overfitting tendency is not found to be markedly stronger for the larger sample sizes.

Information and Other Criteria in Structural Equation Model Selection

Introduction

Since Jöreskog and Sörbom introduced the LISREL model (Jöreskog, 1973, 1977) as well as the program of the same name (Jöreskog & Sörbom, 1976), structural equation modeling (SEM) by means of LISREL and other SEM programs has become very popular in the behavioral sciences. One of the reasons for the popularity is that from the start heavy emphasis was put on model fit assessment. To the χ^2-value with associated p-value (chi-square goodness of fit test) for overall model fit assessment in the previous editions of the LISREL program, the fifth edition (Jöreskog & Sörbom, 1981) added the GFI (goodness of fit index), $AGFI$ (adjusted goodness of fit index), and RMR (root mean square residual). However, all of these early criteria met much criticism. In an influential article by Bentler and Bonett (1980) the criticism on the chi-square goodness of fit test was formulated as follows:

> Large sample theory provides a chi-square goodness-of-fit test for comparing a model against a general alternative model based on correlated variables. This model comparison is insufficient for model evaluation: In large samples virtually any model tends to be rejected as inadequate, and in small samples various competing models, if evaluated, might be equally acceptable. (p. 588)

Bentler and Bonett's article and a long series of subsequent publications (e.g. Bentler, 1990; Bollen, 1986, 1989a; Bozdogan, 1991; Cudeck & Browne, 1983; James, Mulaik & Brett, 1982; Hoelter, 1983; McDonald & Marsh, 1990; Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989) proposed alternatives for the chi-square goodness of fit test and other LISREL fit criteria.

A special class of fit criteria, called information criteria, originated in mathematical statistics outside of the SEM tradition with Akaike (1973). Akaike was himself the first to apply his Akaike Information Criterion (AIC)

to factor analysis models (Akaike, 1987). The class was extended by Bozdogan (1987), Haughton (1988), Haughton and Dudley (1993), Haughton, Haughton, and Izenman (1990), Schwarz (1978) and others. The recent eighth edition of the LISREL program (Jöreskog & Sörbom, 1993a) incorporates many of the new proposals, including two of the information criteria. Up to now, no single criterion has been identified as best by a majority of researchers. It turns out to be impossible to evaluate the quality of the fit criteria on mathematical grounds alone. Even with respect to the information criteria, Bozdogan (1987) observed:

> The preference of one or the other of these criteria in a given situation depends on how "conservative" or "liberal" we want to be in terms of setting the level of significance α per complexity and avoid overfitting and underfitting risks. (p. 368)

To yield more insight into the performance of the criteria, a number of Monte Carlo simulation studies, starting with the Boomsma (1983) study of the LISREL chi-square goodness of fit test, have been carried out in the past. A review of these simulation studies by Gerbing and Anderson (1993) shows that the evaluation of information criteria in comparison to other criteria is rare. An exception is the study of Bandalos (1993), comparing three information criteria with one another and with the cross-validation index (CVI) of Cudeck and Browne (1983), which is theoretically very akin to the information class criteria. This article will present results for a wider variety of criteria from the information as well as the SEM tradition. In view of the growing emphasis on asymptotic behavior considerations (see e.g. McDonald, 1989), we include along with relatively small sample sizes ($n = 100, 400$) considerably larger sample sizes ($n = 1000, 6000$) than found in most other studies.

The criteria are solely understood and applied as selection criteria, to choose the best model according to the criterion out of a set of different models. Some criteria, especially the so-called normed fit measures, have more pretentions, trying to assess in some sense or another the fit of the model on an absolute scale independently from other models. Our evaluation, however, is restricted to the comparative use in the frequently occurring situation that different competing models are fitted to one and the same data set and the criteria printed in the SEM program output are used to identify the best fitting one. The set of competing models will be called selection set. As required by many testing procedures and criteria outside of the information class, the selection set is chosen hierarchically nested (Bentler, 1990): each consecutive

model in the hierarchy is obtained by imposing one or more restrictions on the preceding one.

We start with some general considerations concerning the design of the simulation study, then we explain the information and other selection criteria to be evaluated, and next a detailed description of the simulation study is given, followed by the results and conclusions.

Design of the Simulation Study

A sharp distinction is made between a "true" model and an "analytic" model. A true model has all parameters fixed and is used to generate sample data sets to be analyzed by means of the analytic models in the selection set. An analytic model has one or more free parameters and is either "overfitting" or "underfitting". It is overfitting, if by fixing the free parameters at specific values in the parameter space it is possible to produce the model implied covariance matrix of the true, sample generating model. This is equivalent to the noncentrality parameter of its χ^2-distribution being equal to 0. It is underfitting otherwise (noncentrality parameter greater than 0). The number of free parameters is called the "dimension" of an analytic model. Subtracting the dimension from the number of elements in and below the main diagonal of the covariance matrix gives the "degrees of freedom" of the analytic model. The notion of "correct" analytic model is not to be confused with that of the true (exactly fitting) model. Correct should be understood in a comparative sense. The correct model is, in fact, the most restricted or lowest-dimensional overfitting model in the selection set. By definition the correct model is also the best analytic model in the set, and should ideally be selected by the selection criterion.

Overfitting analytic models with higher dimension than the correct model as well as underfitting analytic models with lower dimension were both included in the selection set to evaluate overfitting as well as underfitting behavior of the selection criteria. A selection criterion is said to overfit, if it tends to select as best an overfitting model of higher dimension than the correct model. Underfitting analytic models were chosen by constraining parameters to be equal which were, in fact, unequal in the true model. Different true models

Figure 1
Examples of selection profiles for five hierarchically nested analytic models: models 1 and 2 overfitting, model 3 correct, models 4 and 5 underfitting

with different degrees of parameter value differences were entertained to evaluate the effects on selection behavior. Underfitting behavior or the tendency to select an analytic model of lower dimension than the correct model, should be avoided more easily for highly separated parameter values (big differences in the true model between the parameter values of constrained parameters) than for only slightly separated parameter values.

The performance of the selection criteria is analyzed in terms of the selection profiles. Figure 1 shows examples. In addition to the ideal selection profile, a well-behaving, a poorly behaving, a moderately underfitting, and a maximally overfitting selection profile are shown. In the ideal selection profile the correct model is selected a 100% of the times. Except for extremely large sample sizes this ideal is too demanding in practice. The well-behaving profile, which combines some underfitting and overfitting behavior with a maximum at the correct model, is a more realistic goal. For slightly separated parameter values (almost equal parameter values in the true model) it can be argued that moderately underfitting behavior is preferable to predominantly overfitting behavior, especially in the case of small sample sizes. Parsimony considerations require that decreasing information as a result of decreasing sample size keeps avoiding overfitting behavior. On the other hand, increasing information as a result of increasing sample size should tend to a well-behaving and in the end to the ideal selection profile. Many fit measures proposed in literature turn out to be useless for selection purposes because they necessarily choose the highest-dimensional overfitting model, shown in the maximally overfitting profile.

Information Criteria

The introduction in 1973 of a model selection criterion by Akaike (Akaike, 1973) initiated much research activity on the behavior and properties of a class of criteria generally known as information criteria. The term "information criterion" arises from the fact that the Kullback-Leibler (K.L.) information distance between the true distribution and its fitted distribution according to a given analytic model is central to the derivation of AIC. The idea of minimizing the K.L. distance in fitting the distribution to the data leads quite naturally to the principle of maximum likelihood. AIC was introduced as an asymptotically unbiased estimator of the mean expected log likelihood of the model (see Sakamoto, Ishiguro, & Kitagawa, 1986; Bozdogan, 1987). The AIC is defined as:

$$AIC(M_i) = \ell_{max}(M_i) - k(M_i) \tag{1}$$

where $\ell_{max}(M_i)$ is the maximum of the log likelihood function for model M_i and $k(M_i)$ is the number of free parameters in this model or its dimension. To select a model according to AIC from a set of m candidate models $\{M_1, M_2, ..., M_m\}$, we choose the model for which AIC is largest. We note that AIC is also often defined as -2 times the AIC defined above; in this case we select the model with the smallest AIC.

The log likelihood function for LISREL model M_i is for n observations:

$$\ell(M_i) = -\frac{n}{2}\log \mid \Sigma_{M_i} \mid -\frac{n}{2}tr(\mathbf{S}\Sigma_{M_i}^{-1}) - \frac{np}{2}\log 2\pi \quad , \tag{2}$$

where Σ_{M_i} is the $p \times p$ model implied covariance matrix and $\mathbf{S}$ is the $p \times p$ sample covariance matrix. The log likelihood function is maximized over the k free parameters in the model. However, instead of maximizing Equation 2, the LISREL program minimizes:

$$\begin{aligned} F(M_i) &= -\frac{2}{n}[\ell(M_i) - \ell_{max}(M_a)] \\ &= \log \mid \Sigma_{M_i} \mid + tr(\mathbf{S}\Sigma_{M_i}^{-1}) - \log \mid \mathbf{S} \mid - p \quad , \end{aligned} \tag{3}$$

where M_a is any general (minimally restricted or saturated) alternative model having $k(M_a) = \frac{1}{2}p(p+1)$ with $\Sigma_{M_a} = \mathbf{S}$ at the maximum of its likelihood

function. Equation 3 differs from Equation 2 only in the negative multiplying factor -2/n and an additive constant (**S** is derived from the data only and thus constant in the log likelihood function). Hence, minimizing Equation 3 leads to the same parameter estimates in model M_i as maximizing Equation 2, while the maximum $\ell_{max}(M_i)$ of Equation 2 is related to the minimum $F_{min}(M_i)$ of Equation 3 as follows:

$$\ell_{max}(M_i) = -\frac{n}{2}F_{min}(M_i) - \frac{n}{2}[\log \mid \mathbf{S} \mid + p(1 + \log 2\pi)] \quad . \tag{4}$$

Note that multiplication of the minimum $F_{min}(M_i)$ by n immediately gives the well-known χ^2-value for testing M_i against a general alternative M_a.[1] In fact, the LISREL program (see Jöreskog & Sörbom, 1993b, p. 119) gives AIC in the following form:

$$\begin{aligned} LAIC(M_i) &= nF_{min}(M_i) + 2k(M_i) \\ &= -2[\ell_{max}(M_i) - \ell_{max}(M_a)] + 2k(M_i) \\ &= -2\,AIC(M_i) + 2\,\ell_{max}(M_a) \quad , \end{aligned} \tag{5}$$

which, however, is seen to be a simple linear function of AIC in Equation 1, because $2\,\ell_{max}(M_a)$ is constant when comparing the models in the selection set. Also Browne and Cudeck's (1993, p. 148) approximation of the expected cross-validation index $ECVI(M_i)$, given by

$$\begin{aligned} ECVI(M_i) \approx\ & F_{min}(M_i) + \frac{2}{n}k(M_i) = \\ & -\frac{2}{n}AIC(M_i) + \frac{2}{n}\ell_{max}(M_a) \quad , \end{aligned} \tag{6}$$

and in contrast to the CVI itself (Cudeck & Browne, 1983) computable on the basis of a single sample, is seen to give the same selection results as AIC,

[1] It should be noted that the values printed in the technical LISREL output during minimization are not F but only $\frac{1}{2}F$, so that the χ^2-value is to be obtained by multiplication by $2n$. It should also be noted that LISREL multiplies in fact by $2(n\text{-}1)$ for getting the χ^2-value. This small difference in case of large samples is due to the fact that LISREL uses the Wishart distribution of the covariance matrix of the multivariate normal observations for modeling.

because also n is constant in comparing the models in the selection set for the same data.

The AIC criterion in Equation 1 is in fact a penalized maximum log likelihood; the penalty function is the number of free parameters or dimension of the model. As such it is an implementation of the principle of parsimony. If we selected a model with the largest ℓ_{max} (smallest F_{min}), we would always select the least restricted or highest-dimensional of several nested models (for example, M_a would necessarily be chosen, if included in the selection set). The selection profiles of ℓ_{max} and F_{min} are a priori known to be the maximally overfitting one in Figure 1. With the introduction of a penalty function in AIC, there is a cost to adding more parameters, with an implied protection against the risk of overfitting.

It is known (e.g. Woodroofe, 1982) that AIC is not asymptotically consistent, in the sense that the probabilities that AIC selects the least-dimensional of two nested overfitting models do not converge to one as the sample size increases to infinity. However, the asymptotic overfitting probability, although greater than zero, is less than one in this and many other situations with overfitting as well as with overfitting and underfitting models in the selection set (Haughton, 1996). McDonald (1989, p. 98), pointing out that in practice all models under consideration are underfitting, suggested that the highest-dimensional model will necessarily be selected by AIC and other information criteria for a sufficiently large sample size. This is true asymptotically in some situations where the selection set contains only underfitting models, but not for practically relevant finite sample sizes up to a size as large as 1000 (Bozdogan & Haughton, 1995). In other situations, even where all models are underfitting, the probability that AIC selects the highest dimensional model does not converge to one (Haughton, Oud, & Jansen, 1996).

The introduction of the BIC criterion by Schwarz in 1978 has led to the class of BIC-type criteria which are asymptotically consistent. The BIC is defined as:

$$BIC(M_i) = \ell_{max}(M_i) - \frac{1}{2}k(M_i)\log n \quad , \tag{7}$$

where again $k(M_i)$ is the number of free parameters in the model and n is the sample size. The model to be selected becomes the one with the largest BIC. The central motivation for BIC is set in a Bayesian context: Given prior probabilities and prior distributions for the unknown parameters for each model, a posterior probability that a given model is the best model can be

formulated, even though, in general, this probability is difficult to calculate directly. By best model we mean again the lowest-dimensional overfitting model. The BIC is known to arise as the leading term in an asymptotic expansion of the posterior probability that a given model is the best one as n goes to infinity (see Schwarz, 1978, and Haughton, 1988). So, in that sense, model selection performed by maximizing BIC is close to a Bayes procedure where one would select a model by maximizing the posterior probability that a model is best.

Using further terms in the asymptotic expansion leads to information criteria as, for instance:

$$BIC^*(M_i) = \ell_{max}(M_i) - \frac{1}{2}k(M_i)\log\frac{n}{2\pi} \quad , \tag{8}$$

(see Haughton, 1988; Haughton et al., 1990), and

$$BICR(M_i) = \ell_{max}(M_i) - \frac{1}{2}k(M_i)\log\frac{n}{2\pi} + \log f[\hat{\theta}(M_i)] + \frac{1}{2}\log det[IFIM_1(M_i)] \quad , \tag{9}$$

where $\hat{\theta}(M_i)$ is the maximum likelihood estimator (M.L.E.) for the vector θ of unknown parameters in model M_i, $IFIM_1$ is the inverse Fisher information matrix for one observation evaluated at the M.L.E. $\hat{\theta}(M_i)$, and f is the density (assumed to be smooth and non-zero) of the prior distribution of the unknown parameter θ on the given model (see Haughton & Dudley, 1993). We note that the $IFIM$ for n observations, equal to $IFIM = (1/n)IFIM_1$, is the estimated asymptotic covariance matrix of the M.L.E. $\hat{\theta}(M_i)$. The $IFIM$ (for n observations) can be extracted from the LISREL output, provided that a sample size of $n+1$ instead of n is specified in the LISREL input (this is due to the fact that LISREL uses the Wishart distribution). We note that $BICR$ can also be written as:

$$BICR(M_i) = \ell_{max}(M_i) + \frac{1}{2}k(M_i)\log 2\pi + \log f[\hat{\theta}(M_i)] + \frac{1}{2}\log det[IFIM(M_i)] \quad , \tag{10}$$

since $det(IFIM_1) = n^{k(M_i)}det(IFIM)$.

From another angle, refining the approximations which lead to AIC and introducing a log n term in the penalty function to achieve consistency, yield a criterion $CAIC$ (Bozdogan, 1987) of the form:

$$CAIC(M_i) = \ell_{max}(M_i) - \frac{1}{2}k(M_i)(1 + \log n) \quad , \tag{11}$$

which is closely related to BIC. $CAIC$ is given in the LISREL program output in the following, linearly related form (see Jöreskog & Sörbom, 1993b, p. 119)

$$\begin{aligned} LCAIC(M_i) &= nF_{min}(M_i) + k(M_i)(1 + \log n) \\ &= -2\,CAIC(M_i) + 2\,\ell_{max}(M_a) \quad . \end{aligned} \tag{12}$$

The criteria BIC, BIC^*, $BICR$, and $CAIC$ are known to be asymptotically consistent (see Haughton, 1989; Woodroofe, 1982; Nishii, 1984, for linear regression; Bozdogan, 1987). In other words, for these criteria, the probabilities of underfitting as well as overfitting converge to zero as the sample size goes to infinity.

In summary, model selection problems in SEM when sample sizes are large are amenable to an information criterion approach, notably since the multivariate normal observations are assumed to be independently identically distributed (i.i.d.). By means of the simulation study, we will compare the performance of the information criteria mentioned above with one another and with fit criteria which are more frequently used in SEM model fit assessment.

Other Criteria

The well-known χ^2-value for testing M_i against M_a,

$$\chi^2(M_i) = -2[\ell_{max}(M_i) - \ell_{max}(M_a)] = nF_{min}(M_i) \quad , \tag{13}$$

is as such not useful as a selection criterion, because, like ℓ_{max} and F_{min} with both of which it is linearly related, it necessarily selects the least restrictive

model (the model M_i for which k is largest) and thus has as its selection profile the maximally overfitting one in Figure 1. However, via the degrees of freedom of the model

$$v(M_i) = \frac{1}{2}p(p+1) - k(M_i) \quad , \tag{14}$$

the associated p-value

$$p(M_i) = Pr[\chi^2_v > \chi^2(M_i)] \quad , \tag{15}$$

accounts for the dimension k of the model by referring to the χ^2-distribution with v degrees of freedom: the higher the dimension (the lower the degrees of freedom) is for the same χ^2-value, the lower the fit as measured by the p-value. The criticism on the χ^2 p-value especially concerns its use as a testing criterion (rejecting the model, if p is smaller than the α-level, accepting otherwise). As a testing device its power in rejecting underfitting models (noncentrality parameter greater than 0), although increasing for decreasing dimension, heavily depends on the α-level and the sample size chosen, while the test is constructed such that overfitting models (central χ^2) have equal acceptance probabilities for a constant α. Formally, χ^2-testing is not a selection device, because depending on the α-level chosen all nonsaturated models could be rejected or all could be accepted. Using the p-value as a selection criterion avoids the dependence of the result on the chosen α-level and provides a good standard for evaluating the behavior of other fit criteria. As these have almost all been constructed as alternatives to χ^2-testing, they may be expected to improve upon the selection behavior of the χ^2 p-value.

The p-value will be included as a selection criterion in the simulation study together with the other early LISREL fit criterion, the adjusted goodness of fit criterion $AGFI$:

$$AGFI(M_i) = 1 - \frac{\frac{1}{2}p(p+1)}{v(M_i)}[1 - GFI(M_i)] \quad , \tag{16}$$

which is based on the goodness of fit criterion GFI

$$GFI(M_i) = 1 - \frac{tr[(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}_{M_i} - \mathbf{I})^2]}{tr[(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}_{M_i})^2]} \quad . \tag{17}$$

GFI and $AGFI$ are formulated in analogy to the coefficient of determination and the correction for bias of a squared multiple correlation, respectively. $\hat{\Sigma}_{M_i}$ in GFI is the $p \times p$ model implied covariance matrix evaluated at the M.L.E. $\hat{\theta}(M_i)$ and $\mathbf{I}$ is the $p \times p$ identity matrix. Both GFI and $AGFI$ reach their maximum 1 only for perfect fit, that is $\hat{\Sigma}_{M_i} = \mathbf{S}$. In contrast to $AGFI$, however, which accounts for the model dimension k, GFI is just as meaningless for selection purposes as the χ^2-value. GFI is a monotone transformation of χ^2 for all practically relevant true models (Maiti & Mukherjee, 1990, p. 722) and so a priori known again to give the maximally overfitting selection profile (see Figure 1).

Bentler and Bonett (1980) and Bentler (1990) proposed the normed fit index NFI,

$$NFI(M_i) = \frac{\ell_{max}(M_i) - \ell_{max}(M_b)}{\ell_{max}(M_a) - \ell_{max}(M_b)} = \frac{\chi^2(M_b) - \chi^2(M_i)}{\chi^2(M_b)} \quad , \tag{18}$$

where model M_b, called the "baseline" or "null-model", is a more restricted model than each of the models M_i in the selection set $\{M_1, M_2, ..., M_m\}$ and M_a is any saturated model as in Equation 3. In addition, Bentler and Bonett require the sequence of models M_a, M_1, $M_2, ..., M_m$, M_b to be a hierarchically nested one: each consecutive model to the right is obtained by imposing one or more restrictions on the preceding model to the left. Although NFI is indeed normed in the sense that each model M_i in the selection set gets a goodness of fit measure on a scale from 0 to 1, it does not improve on the χ^2-value or GFI as a selection criterion. Like the χ^2-value and GFI, NFI selects the least restrictive model in the selection set as best, because $\ell_{max}(M_a)$ and $\ell_{max}(M_b)$ as well as $\chi^2(M_b)$ are constants in comparing different models in the selection set. A series of extensions of Bentler and Bonett's NFI have been formulated, however, which are all based on $\chi^2(M_i)$ and $\chi^2(M_b)$ and associated degrees of of freedom, but penalize in different ways through $v(M_i)$ for the model dimension $k(M_i)$ and could therefore be useful selection criteria. These include the nonnormed fit index $NNFI$ (Bentler & Bonett, 1980; Bentler, 1990), ρ_1 (Bollen, 1986), parsimony fit index PFI (James, Mulaik, & Brett, 1982), incremental fit index IFI (Bollen, 1989a), new nonnormed fit index FI (Bentler, 1990), and new normed or comparative fit index CFI (Bentler, 1990):

$$NNFI(M_i) = \frac{\chi^2(M_b)/v(M_b) - \chi^2(M_i)/v(M_i)}{\chi^2(M_b)/v(M_b) - 1} \quad , \tag{19}$$

$$\rho_1(M_i) = \frac{\chi^2(M_b)/v(M_b) - \chi^2(M_i)/v(M_i)}{\chi^2(M_b)/v(M_b)} \quad , \tag{20}$$

$$PFI(M_i) = \frac{v(M_i)}{v(M_b)} \left[\frac{\chi^2(M_b) - \chi^2(M_i)}{\chi^2(M_b)}\right] \quad , \tag{21}$$

$$IFI(M_i) = \frac{\chi^2(M_b) - \chi^2(M_i)}{\chi^2(M_b) - v(M_i)} \quad , \tag{22}$$

$$FI(M_i) = 1 - \frac{\chi^2(M_i) - v(M_i)}{\chi^2(M_b) - v(M_b)} \quad , \tag{23}$$

$$\begin{aligned} CFI(M_i) &= 1 - \frac{\lambda(M_i)}{\lambda(M_b)} \quad \text{with} \\ \lambda(M_i) &= max[\chi^2(M_i) - v(M_i),\ 0] \quad \text{and} \\ \lambda(M_b) &= max[\chi^2(M_b) - v(M_b),\ \lambda(M_i)] \quad . \end{aligned} \tag{24}$$

Bentler and Bonett's $NNFI$ was originally introduced by Tucker and Lewis (1973) in the context of exploratory factor analysis. $\lambda(M_i)$ and $\lambda(M_b)$ in Bentler's CFI are motivated by the fact that $E[\chi^2(M_i) - v(M_i)] \geq 0$ and $E[\chi^2(M_b) - v(M_b)] \geq E[\chi^2(M_i) - v(M_i)]$. Thus CFI avoids the possibility in FI that for finite samples $[\chi^2(M_i) - v(M_i)]/[\chi^2(M_b) - v(M_b)]$ becomes negative or exceeds 1. Note that $NNFI$ and ρ_1 are equivalent as model selection criteria since $\chi^2(M_b)/v(M_b)$ is constant.

It is important to note that the choice of the baseline model does not affect the selection behavior of the fit measures involved except that of the parsimony fit index PFI (Equation 21). The reason is that $\chi^2(M_b)$ and $v(M_b)$ enter the equations only linearly ($NNFI, \rho_1, FI, CFI$) or monotonely (IFI). It is because of the term $v(M_i)\chi^2(M_b)$ in Equation 21 that the selection behavior

of PFI depends on the choice of the baseline model. Because for $PFI(M_j) = PFI(M_i)$ one derives

$$\chi^2(M_b) = \frac{v(M_j)\chi^2(M_j) - v(M_i)\chi^2(M_i)}{v(M_j) - v(M_i)} \quad ,$$

assuming M_j more restricted or parsimonious than M_i, $PFI(M_j) < PFI(M_i)$ implies that $\chi^2(M_b)$ must be in the range

$$\chi^2(M_j) < \chi^2(M_b) < \frac{v(M_j)\chi^2(M_j) - v(M_i)\chi^2(M_i)}{v(M_j) - v(M_i)} \quad .$$

The range is the more limited the smaller the χ^2 difference between the models compared and the larger the difference in their degrees of freedom. For all baseline models with $\chi^2(M_b)$-values outside of this range the more parsimonious model M_j necessarily fits best according to PFI. Replacing the baseline model by one inside or outside of this range changes the PFI selection behavior from $PFI(M_j) > PFI(M_i)$ to $PFI(M_j) < PFI(M_i)$ or vice versa.

The class of fit indices started by Bentler and Bonett with NFI and $NNFI$ has been extended with many more members under a variety of names. However, just as ρ_1 is monotonely related to $NNFI$, most of the other proposals turn out to be monotonely or linearly related or even equal to the indices in Equations 19 through 24 and therefore do not lead to new selection criteria in the sense taken here. For example, McDonald and Marsh (1990) mention the indices c_k and h_k. The first one is equal to IFI (Equation 22), attributed by Bentler (1990) to Bollen (1989a) who called it Δ_2 (see also Bollen, 1990). The second one is equal to FI (Equation 23) and seems to have been invented by Bentler (1990, p. 250) and by McDonald and Marsh (1990, p. 243) independently.

Two original contributions, that will also be included in the simulation study, are Hoelter's (1983) critical-n index CN and Cudeck and Browne's (1983) cross-validation index CVI.

$$CN(M_i) = \frac{\text{crit } \chi^2_{v,\alpha}}{F_{min}(M_i)} + 1 \quad . \tag{25}$$

Here crit $\chi^2_{v,\alpha}$ is the critical χ^2-value for degrees of freedom $v(M_i)$ and chosen significance level α (i.e. the $100(1-\alpha)$th percentile for the central χ^2 distribution with $v(M_i)$ degrees of freedom). For a well fitting model (small F_{min})

the n making $\chi^2(M_i)$ significant will be large; for a badly fitting model n will be small. Like the p-value CN penalizes for the dimension k by referring to the χ^2-distribution with v degrees of freedom.

Finally, the cross-validation index CVI,

$$\begin{aligned} CVI(M_i) &= F(\hat{\Sigma}_{M_i,A}, \mathbf{S}_B) \\ &= \log \mid \hat{\Sigma}_{M_i,A} \mid + tr(\mathbf{S}_B \hat{\Sigma}_{M_i,A}^{-1}) - \log \mid \mathbf{S}_B \mid - p \ , \end{aligned} \tag{26}$$

is in the form of the LISREL fitting function F (see Equation 3), but it avoids selection of the least restricted model by inserting for $\mathbf{S}$ the covariance matrix $\mathbf{S}_B$ of a sample B (validation sample) that is different from the sample A (calibration sample) used for the maximum likelihood estimation of the model M_i and the computation of $\hat{\Sigma}_{M_i,A}$. It should be noted that it is assumed $n_A = n_B$ and that no minimization is involved in the computation of CVI. $F(\hat{\Sigma}_{M_i,A}, \mathbf{S}_B)$ is merely a discrepancy measure between the validation-sample covariance matrix and the calibration-sample model implied covariance matrix.

Simulation Study

Consider the following confirmatory factor analysis model:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & \lambda_{53} \\ 0 & 0 & \lambda_{63} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{pmatrix}$$

The vectors x and ξ are assumed to have zero expectation. The measurement errors $\delta_1, \delta_2, \cdots, \delta_6$ are assumed to be independently and identically distributed according to $N(0, \sigma_{\delta_i}^2)$, $i = 1, \cdots, 6$. We also assume that ξ and δ are uncorrelated, and that the covariance matrix $\mathbf{\Phi}$ of ξ has ones on its diagonal.

In this study we will use three true models, differing in the true values of the parameters $\sigma_{\delta_i}^2$ in the model above:

True model I:

$\lambda_{11} = \lambda_{21} = 1, \lambda_{32} = \lambda_{42} = 2,\ \lambda_{53} = \lambda_{63} = 3,$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = .25,\ \sigma^2_{\delta_3} = \sigma^2_{\delta_4} = .50,\ \sigma^2_{\delta_5} = \sigma^2_{\delta_6} = .75,$
$\mathbf{\Phi} = \mathbf{I}$, the three dimensional identity matrix.

True model II:

$\lambda_{11} = \lambda_{21} = 1,\ \lambda_{32} = \lambda_{42} = 2,\ \lambda_{53} = \lambda_{63} = 3,$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = .35,\ \sigma^2_{\delta_3} = \sigma^2_{\delta_4} = .50,\ \sigma^2_{\delta_5} = \sigma^2_{\delta_6} = .65,$
$\mathbf{\Phi} = \mathbf{I}$, the three dimensional identity matrix.

True model III:

$\lambda_{11} = \lambda_{21} = 1,\ \lambda_{32} = \lambda_{42} = 2,\ \lambda_{53} = \lambda_{63} = 3,$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = .45,\ \sigma^2_{\delta_3} = \sigma^2_{\delta_4} = .50,\ \sigma^2_{\delta_5} = \sigma^2_{\delta_6} = .55,$
$\mathbf{\Phi} = \mathbf{I}$, the three dimensional identity matrix.

True model I shows relatively large differences in measurement error variances (.25, .50, .75), model II intermediate differences (.35, .50, .65), and model III very small differences (.45, .50, .55).

Given data generated according to one of the three models above, we entertain the following hierarchically nested analytic models:

Overfitting models:

Analytic model 1 (dimension 12):

$\lambda_{11} = \lambda_{21},\ \lambda_{32} = \lambda_{42},\ \lambda_{53} = \lambda_{63},$
All $\sigma^2_{\delta_i}$ are estimated freely,
$\mathbf{\Phi}$ has ones on its diagonal, but the off-diagonal elements ϕ_{12}, ϕ_{23}, ϕ_{13} are estimated freely.

Analytic model 2 (dimension 9):

$\lambda_{11} = \lambda_{21},\ \lambda_{32} = \lambda_{42},\ \lambda_{53} = \lambda_{63},$
All $\sigma^2_{\delta_i}$ are estimated freely,
$\mathbf{\Phi} = \mathbf{I}$.

Correct model:

Analytic model 3 (dimension 6):

$\lambda_{11} = \lambda_{21},\ \lambda_{32} = \lambda_{42},\ \lambda_{53} = \lambda_{63},$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2},\ \sigma^2_{\delta_3} = \sigma^2_{\delta_4},\ \sigma^2_{\delta_5} = \sigma^2_{\delta_6},$
$\mathbf{\Phi} = \mathbf{I}$.

Underfitting models:

Analytic model 4 (dimension 5):

$\lambda_{11} = \lambda_{21},\ \lambda_{32} = \lambda_{42},\ \lambda_{53} = \lambda_{63},$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = \sigma^2_{\delta_3} = \sigma^2_{\delta_4},\ \sigma^2_{\delta_5} = \sigma^2_{\delta_6},$
$\mathbf{\Phi} = \mathbf{I}.$

Analytic model 5 (dimension 4):

$\lambda_{11} = \lambda_{21},\ \lambda_{32} = \lambda_{42},\ \lambda_{53} = \lambda_{63},$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = \sigma^2_{\delta_3} = \sigma^2_{\delta_4} = \sigma^2_{\delta_5} = \sigma^2_{\delta_6},$
$\mathbf{\Phi} = \mathbf{I}.$

"Correct" for analytic model 3 should be understood in the comparative sense of the most restricted overfitting model. Evidently, there is only one completely correct analytic model: the model with all parameters fixed and equal to the values of the true model. The analytic models 1-5 are designed to help us investigate how well the information and other criteria perform at accepting the "true" restrictions (only "true" restrictions are present in analytic models 1-3 but most in model 3 and fewest in model 1) and rejecting the "false" restrictions (present in model 5 and to a lesser degree in model 4).

A last and still more restrictive underfitting model than model 5, to be used as the baseline model M_b in the Bentler-type criteria ("the most restrictive, theoretically defensible model"; Bentler & Bonett, 1980) closes the hierarchy:

Model b (dimension 1):

$\lambda_{11} = \lambda_{21} = \lambda_{32} = \lambda_{42} = \lambda_{53} = \lambda_{63} = 0,$
$\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = \sigma^2_{\delta_3} = \sigma^2_{\delta_4} = \sigma^2_{\delta_5} = \sigma^2_{\delta_6},$
$\mathbf{\Phi} = \mathbf{I}.$

Because the most underfitting analytic model 5 (dimension 4) has only four free parameters (the three factor loadings and the single measurement error variance), leading to pairwise equal variance estimates for the observed variables, introducing additional uncorrelatedness between the observed variables by eliminating factor loadings leads to a single, common variance estimate for all observed variables. So, the baseline model becomes the true counterpart of the saturated model in representing the observed variables as independent identically distributed variables. Going up the hierarchy, factor loadings are added and successively more constraints on the measurement error variances are freed until in the most overfitting model 1 (dimension 12) additionally the three factor covariances are freed.

The procedure of the simulation study was as follows:

1. The model implied covariance matrices $\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}_\delta$ were computed for the true models I, II, and III by means of the LISREL program (Jöreskog & Sörbom, 1993a), fixing all parameters at the true model values.

2. The LISREL program was used to compute the lower triangular matrices $\mathbf{T}$ such that $\mathbf{\Sigma} = \mathbf{TT}'$ (see the procedure in Jöreskog & Sörbom, 1994, pp. 7-8) for the true models I, II, and III.

3. The PRELIS program (Jöreskog & Sörbom, 1994, pp. 8-10) was used to generate four sets (sample sizes $100, 400, 1000, 6000$) of 500 sample covariance matrices of six multinormally distributed variables on the basis of $\mathbf{T}$. This was repeated for the true models I, II, and III.

4. Each sample covariance matrix in each of the 12 sets of 500 covariance matrices was analyzed by means of the five analytic models 1, 2, 3, 4, 5, using the LISREL program. As starting values the parameter values of the corresponding true model I, II, or III were used. The LISREL print output files, goodness of fit matrices GF output files, asymptotic covariance matrices of estimated parameters EC output files, parameter matrices LX, PH, and TD output files were collected for each set of 500 samples.

5. The LISREL output files were checked for Heywood cases (negative estimates of measurement error variances) and other abnormal results.

6. The values of the following 15 information and other criteria were taken from the LISREL output or computed on the basis of information from the LISREL output according to the equations given below between parentheses. For the computation of the cross-validation index CVI a different sample was used for validation, so that each sample was used once as calibration sample and once as validation sample.

   1. AIC (Equation 1)
   2. BIC (Equation 7)
   3. BIC^* (Equation 8)
   4. $BICR$ (Equation 10)
   5. $CAIC$ (Equation 11)
   6. CVI (Equation 26)
   7. p (Equation 15)
   8. $AGFI$ (Equation 16)

9. $NNFI$ (Equation 19)
10. PFI (Equation 21)
11. IFI (Equation 22)
12. FI (Equation 23)
13. CFI (Equation 24)
14. $CN(\alpha = .05)$ (Equation 25)
15. $CN(\alpha = .01)$ (Equation 25)

7. The percentages within each set of 500 samples were computed that each of the five analytic models was chosen as best by each of the 15 criteria. Percentages were computed after elimination of the Heywood cases, while ties (exactly equal criterion values for different analytic models) were equally divided over the analytic models involved.

Results and Conclusions

The results of the model selection procedure are given as percentages in Table 1 and, based on this table, graphically in the form of selection profiles in Figure 2. All 30000 LISREL runs were completed without minimization problems and no abnormal results were encountered other than Heywood cases. Apart from one sample of $n = 400$ generated by true model I and analyzed by analytic model 1, Heywood cases occurred only for samples of the smallest size $n = 100$ analyzed by the highest-dimensional analytic models 1 and 2. Under true model I, II, and III, respectively, 40, 34, and 47 Heywood cases were produced by analytic model 1, and 34, 38, and 47 by analytic model 2. Heywood cases were left out of the computations.

Ties occurred only for CFI and $AGFI$. Ties for CFI are not avoidable but a logical consequence of its definition (Equation 24). In fact, Bentler's new normed or comparative fit index CFI is a tie generating adjustment of his new nonnormed fit index FI, making all (typically different) FI values below 0 equal to 0 and above 1 equal to 1. When these ties are divided equally over the models involved as in Table 1 and Figure 2, the adjustment is clearly seen in almost all cases to lower the percentage of correct model selection and to increase the percentages of choices of overfitting models. For $AGFI$ the percentage part of ties in the selection percentages involved is negligible (never exceeding 1.2%), always concerns values close to the upper limit 1, and

Table 1: Percentages of best model selection in the set of five hierarchically nested analytic models (1 and 2 overfitting, 3 correct, 4 and 5 underfitting) under true models I, II and III, and sample sizes n=100, 400, 1000, 6000

True model I

N		100	400	1000	6000
AIC	1	4.6	5.4	4.2	3.4
	2	7.8	8.8	8.8	10.4
	3	86.2	85.8	87.0	86.2
	4	1.4			
	5				
BIC	1				
	2	0.4			
	3	84.6	100.0	100.0	100.0
	4	15.0			
	5				
BIC*	1	0.2			
	2	3.8	0.6	0.2	
	3	93.2	99.4	99.8	100.0
	4	2.8			
	5				
BICR	1				
	2	13.0	2.4	0.4	
	3	72.8	97.6	99.6	100.0
	4	14.2			
	5				
CAIC	1				
	2				
	3	81.8	100.0	100.0	100.0
	4	17.8			
	5	0.4			
CVI	1	8.0	30.8	16.4	8.0
	2	13.4	12.6	15.4	18.8
	3	72.2	56.6	68.2	73.2
	4	6.4			
	5				
p	1	27.8	33.8	34.6	33.0
	2	24.8	24.0	24.6	24.8
	3	47.2	42.2	40.8	42.2
	4	0.2			
	5				
AGFI	1	29.6	35.6	41.2	36.2
	2	27.6	23.7	22.2	24.4
	3	42.6	40.7	36.6	39.4
	4	0.2			
	5				

N		100	400	1000	6000
NNFI	1	32.4	35.8	41.0	36.4
	2	19.0	23.2	22.6	24.4
	3	47.8	41.0	36.4	39.2
	4	0.8			
	5				
PFI	1				
	2				
	3	0.8			
	4	19.6	3.0		
	5	79.6	97.0	100.0	100.0
IFI	1	25.6	32.6	29.2	29.6
	2	18.0	24.0	25.0	26.0
	3	56.2	43.4	45.8	44.4
	4	0.2			
	5				
FI	1	26.0	32.4	29.2	29.6
	2	17.6	24.2	25.0	26.0
	3	56.2	43.4	45.8	44.4
	4	0.2			
	5				
CFI	1	31.9	39.6	36.2	36.9
	2	26.6	27.3	29.7	29.9
	3	39.4	33.0	34.0	33.2
	4	2.2			
	5				
CN(.05)	1	43.0	50.2	51.0	49.0
	2	23.2	24.4	24.4	23.8
	3	33.2	25.4	24.6	27.2
	4	0.6			
	5				
CN(.01)	1	48.0	57.6	57.0	55.0
	2	24.6	21.2	22.6	23.0
	3	27.4	21.2	20.4	22.0
	4				
	5				

Table 1: (*continued*)
True model II

N		100	400	1000	6000
AIC	1	2.6	3.2	5.0	4.2
	2	8.2	8.2	8.0	10.8
	3	52.8	86.8	87.0	85.0
	4	31.6	1.8		
	5	4.8			
BIC	1				
	2				
	3	28.2	85.0	99.8	100.0
	4	43.0	15.0	0.2	
	5	28.8			
BIC*	1	0.8			
	2	1.8			
	3	50.8	92.0	100.0	100.0
	4	34.8	8.0		
	5	11.8			
BICR	1				
	2	10.8	0.4	0.4	
	3	25.8	85.2	99.4	100.0
	4	38.6	14.4	0.2	
	5	24.8			
CAIC	1				
	2				
	3	20.6	81.6	99.4	100.0
	4	39.6	18.4	0.6	
	5	39.8			
CVI	1	5.8	26.4	18.6	11.2
	2	12.0	14.4	13.4	19.4
	3	50.8	54.0	67.4	69.4
	4	23.0	4.8	0.6	
	5	8.4	0.4		
p	1	24.6	29.4	26.6	37.4
	2	27.2	27.2	29.0	21.2
	3	36.4	42.6	44.4	41.4
	4	10.8	0.8		
	5	1.0			
AGFI	1	27.2	32.2	33.0	39.7
	2	26.0	24.2	28.2	21.7
	3	34.6	43.0	38.8	38.6
	4	10.4	0.6		
	5	1.8			

N		100	400	1000	6000
NNFI	1	31.2	31.8	33.2	39.8
	2	18.4	24.6	28.2	21.6
	3	36.6	42.8	38.6	38.6
	4	12.8	0.8		
	5	1.0			
PFI	1				
	2				
	3				
	4	1.2			
	5	98.8	100.0	100.0	100.0
IFI	1	24.0	29.4	25.0	34.2
	2	18.0	26.8	24.2	21.6
	3	43.8	43.0	50.8	44.2
	4	13.0	0.8		
	5	1.2			
FI	1	23.4	29.4	25.0	34.2
	2	18.6	26.8	24.2	21.6
	3	43.8	43.0	50.8	44.2
	4	13.0	0.8		
	5	1.2			
CFI	1	26.9	34.3	35.2	40.0
	2	22.2	28.2	31.7	28.4
	3	36.7	35.7	33.1	31.6
	4	11.3	1.8		
	5	2.9			
CN(.05)	1	40.2	44.0	49.4	52.0
	2	24.0	24.0	25.6	23.0
	3	28.4	31.2	25.0	25.0
	4	7.4	0.8		
	5				
CN(.01)	1	48.0	50.8	55.8	59.4
	2	20.6	23.4	23.2	20.8
	3	24.6	25.2	21.0	19.8
	4	6.8	0.6		
	5				

Table 1: (*continued*)
True model III

N		100	400	1000	6000
AIC	1	1.6	1.6	2.6	3.8
	2	6.0	5.0	9.4	11.6
	3	11.4	33.4	49.4	84.0
	4	18.0	33.8	33.8	0.6
	5	63.0	26.2	4.8	
BIC	1				
	2				
	3	1.8	2.6	13.4	89.2
	4	9.8	21.2	43.4	10.8
	5	88.4	76.2	43.2	
BIC*	1				
	2	1.6			
	3	7.8	9.0	25.2	94.0
	4	14.8	34.0	47.8	6.0
	5	75.8	57.0	27.0	
BICR	1				
	2	9.8	0.2		0.2
	3	1.6	2.6	14.6	90.0
	4	10.0	23.2	43.6	9.8
	5	78.6	74.0	41.8	
CAIC	1				
	2				
	3	0.4	1.8	8.8	87.2
	4	6.2	15.0	38.2	12.8
	5	93.4	83.2	53.0	
CVI	1	4.6	28.8	16.4	7.6
	2	9.8	7.8	10.2	17.2
	3	21.2	27.0	46.4	71.2
	4	26.2	19.8	21.0	4.0
	5	38.2	16.6	6.0	
p	1	20.0	28.6	28.6	27.6
	2	26.6	24.4	24.2	29.0
	3	16.0	21.4	35.6	43.4
	4	14.0	17.0	11.0	
	5	23.4	8.6	0.6	
AGFI	1	22.4	34.3	32.0	32.9
	2	27.0	22.8	22.4	25.5
	3	13.6	19.7	34.8	41.6
	4	12.4	15.2	9.8	
	5	24.6	8.0	1.0	

N		100	400	1000	6000
NNFI	1	24.4	34.2	32.2	32.8
	2	20.4	22.4	22.6	25.2
	3	17.0	20.0	34.2	42.0
	4	14.2	15.2	10.0	
	5	24.0	8.2	1.0	
PFI	1				
	2				
	3				
	4				
	5	100.0	100.0	100.0	100.0
IFI	1	20.6	23.6	28.4	27.0
	2	11.4	21.2	22.0	26.2
	3	22.4	26.0	35.0	46.8
	4	19.2	21.0	13.8	
	5	26.4	8.2	0.8	
FI	1	20.6	23.6	28.4	27.0
	2	12.2	21.2	22.0	26.2
	3	22.0	26.0	35.0	46.8
	4	18.8	21.0	13.8	
	5	26.4	8.2	0.8	
CFI	1	24.5	29.0	31.7	35.8
	2	17.0	23.5	26.6	30.4
	3	21.1	20.7	28.3	33.5
	4	17.5	17.3	10.7	0.4
	5	19.8	9.5	2.7	
CN(.05)	1	36.8	47.0	44.6	49.0
	2	23.6	23.2	26.0	25.8
	3	14.0	14.4	23.8	25.2
	4	11.8	9.2	5.2	
	5	13.8	6.2	0.4	
CN(.01)	1	43.8	53.2	52.2	54.2
	2	22.0	23.6	25.2	24.2
	3	12.8	12.4	19.0	21.6
	4	11.8	6.2	3.2	
	5	9.6	4.6	0.4	

Figure 2: Selection profiles for five hierarchically nested models (1 and 2 overfitting, 3 correct, 4 and 5 underfitting) under true models I, II and III, and sample sizes n=100, 400, 1000, 6000, based on the percentages in Table 1

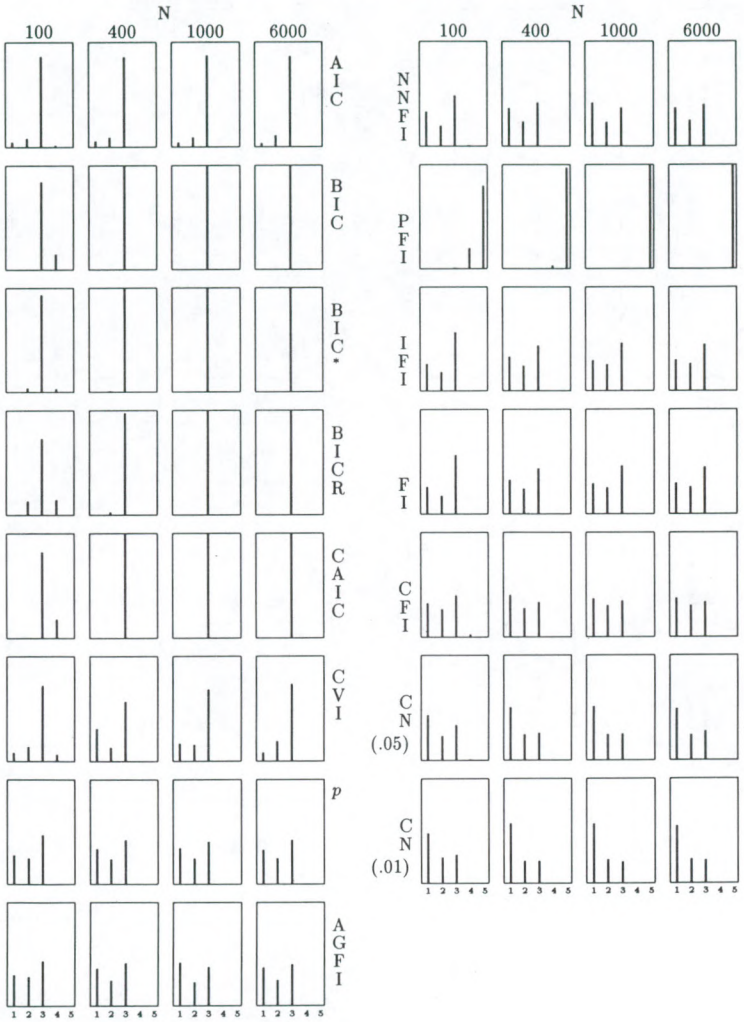

Figure 2: (*continued*)

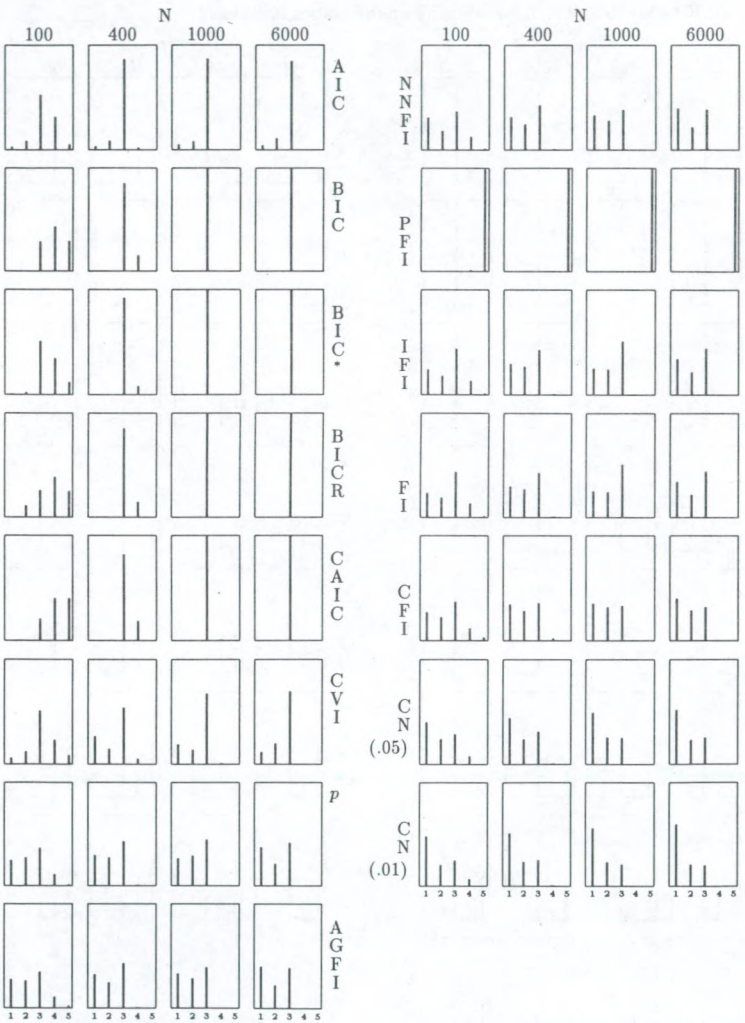

Figure 2: (*continued*)

is very likely to be caused by the fact that differences could not show up in the five decimals given by the LISREL program.

Recalling that the correct model is the six-dimensional model 3, a quick glance at the percentages of correct model selections in Table 1 and Figure 2 for all the criteria indicates that overall the information criteria obtain higher percentages than other criteria, at least for true models I and II. This is very clear for true model I where the three measurement error variances are most separated (.25, .50, .75), and holds almost always for true model II (medium separation of the measurement error variances: .35, .50, .65) with some exceptions for the lowest sample size $n = 100$. In the case of true model III (extremely low separation of the measurement error variances: .45, .50, .55), the information criteria perform very well for the largest sample size $n = 6000$, but increasingly underfit for lower sample sizes. Here the behavior of information criteria should be viewed as a desirable application of Occam's razor: do not make the model more complex than warranted by the data. Relatively high percentages of correct model selections occur here for the cross-validation index CVI, which in comparison to the criteria outside of the information class performs also very well in the cases of true models I and II. It should be noted that CVI requires in addition to the calibration sample a validation sample and is thus based on twice the amount of data.

The known tendency of AIC to overfit is clearly visible in Table 1 and Figure 2, when AIC is compared with the often ideal selection profiles of the other information criteria. For AIC the probability of overfitting is known to not approach zero as the sample size increases (non-consistency), and we indeed see that the percentages of overfitting selections, although small, are persistent for the larger sample sizes. However, in contrast to the suggestions in the literature, there is no marked increase for increasing sample size. Even for a sample size as large as $n = 6000$ the overfitting percentage (analytic models 1 and 2 combined) for AIC never exceeds 16%.

The best performing criterion in the information class seems to be BIC^*, which for the low sample sizes is somewhat less underfitting than the other consistent information criteria. Somewhat stronger underfitting behavior is shown by $CAIC$. The penalty function for BIC^* is larger than that of AIC but smaller than that of BIC, so it seems to provide with a satisfactory compromise.

One might have expected $BICR$ to perform better than BIC^*, since it contains further terms in the expansion of the log of the posterior probability that a model is best (see Equations 9 and 8); however a difficulty in using

$BICR$ is to select a suitable prior distribution for the unknown parameters. A prior with non-zero density on the parameter space is needed for each model. For our calculations, we used the Cauchy distribution on the real line for the λ's (with density $1/[\pi(1+\lambda^2)]$); its tails are thicker than those of the normal distribution. For the σ^2's, we used a Cauchy distribution on the positive half-line, with density $f(\sigma^2) = 2/[\pi(1+(\sigma^2)^2])$, and for the non-diagonal ϕ's in the covariance matrix $\mathbf{\Phi}$ of the latent factor ξ, we used a uniform distribution on $[-1,1]$ (the non-diagonal ϕ's are restricted to $[-1,1]$ since the diagonal ϕ's are assumed to be equal to one). Other prior distributions are possible, such as Jeffreys' non informative prior (see, e.g., Kass, 1989), and it is conceivable that $BICR$ might perform better with a different choice of prior. But the main focus of this article is on comparing the information criteria with the more classical fit indices, so we do not pursue other priors here.

Among the other criteria, the CVI (Cudeck and Browne, 1983) gives by far the best results. The approach consisting in evaluating the performance of the model on an independent data set is an interesting one, used in many applications, and works quite well. In fact Cudeck and Browne (1983) did calculate AIC and BIC along with CVI but not in the context of evaluating the performance of criteria by the frequency of selection of the correct model. Our results indicate that CVI selects the overfitting models more frequently than AIC, but selects the underfitting models less frequently than BIC. This confirms one of the observations of Cudeck and Browne that BIC appears more conservative than CVI by choosing less frequently higher dimensional models. Browne and Cudeck (1989, 1993) also developed the expected cross validation $ECVI$, which is based on a single sample, but as pointed out by them and Jöreskog (1993) and shown in Equation 6, it gives the same rank order between competing models as AIC and is thus equivalent to AIC as a model selection criterion.

All remaining criteria - the χ^2 p-value, $AGFI$, the five Bentler-type measures ($NNFI$, PFI, IFI, FI, CFI), and the two critical-n indices - have overall considerably lower correct model selection percentages. The best performers in this group are the almost identically behaving IFI and FI. With the exception of PFI, all these remaining criteria exhibit strong overfitting tendencies which do not seem to decrease as the sample size gets larger. It has been argued in the literature that the statistic $nF_{min}(M_i)$ in Equation 13 may not be asympotically distributed as a central χ^2 (see e.g. Bollen, 1989b, pp. 266-268; Bozdogan, 1991). The asymptotic distribution will certainly fail to be a central χ^2 if the model M_i in Equation 13 is not overfitting (does not cover the true parameter values of the distribution generating the data). Pre-

vious literature has therefore pointed to the problem of model misspecification as a limitation to χ^2 based criteria, and has also stated that when the sample size is large, models are likely to be rejected for trivial reasons. The results of this study indicate the presence of problems with χ^2 based criteria even when there is no misspecification, as is the case here. The built-in dimension penalizations simply turn out not to work properly: the less restrictive analytic model 1 is more frequently chosen than the more restrictive analytic model 2. It is interesting in this respect that the heavily criticized p-value of the χ^2 test, which typically motivated the construction of the alternative measures, shows in many cases as good or better results with often less overfitting.

The overfitting tendency is particularly strong in both CN measures. However, because the choice of $\alpha = .05$ instead of $\alpha = .01$ systematically decreases the percentage for overfitting and increases the percentage for correct model 3, is seems worthwile to investigate whether choosing still higher α-levels than the customary $\alpha = .05$ would improve the performance of the CN criterion.

While most of the poorly performing criteria show a clear tendency to overfit, the problem with the parsimony fit index CFI is just the opposite: it is parsimonious indeed, as it underfits in an extreme way in all true model and sample size situations. Mulaik et al. (1989, pp. 436-437) state that PFI "has certain affinities to the Akaike (1987) AIC lack of fit index, which also penalizes a model for losses in degrees of freedom". There can be no doubt that in contrast to AIC, PFI is an example of extreme overpenalization. It penalizes not only for the loss in degrees of freedom but also heavily for the baseline model chosen.

In summary:

- In this study where the data were generated by a model (called the "true" model) that was covered by three of five hierarchically ordered analytic models and most restrictively by the so-called "correct" analytic model, the information criteria were overall more successful at the task of selecting the correct analytic model than the other criteria considered. However, the two sample cross-validation criterion remains an attractive option.

- Of the information criteria considered, the best performer seems to be BIC^* which is somewhat less overfitting than AIC, and somewhat less underfitting than the other information criteria (BIC, $BICR$, $CAIC$).

- All other criteria included in the study, with the exception of the extremely underfitting parsimony fit index PFI, showed overfitting problems and did not or hardly improve on the well-known χ^2 p-value as a selection

criterion.

Discussion

A problem inherent to simulation studies as the one reported here, is that generalization of the results to true models and sample sizes outside of the ones used in the study is not warranted. However, the simulation procedure by means of the PRELIS and LISREL programs, explained in this article, is easily repeatable for other kinds of models and sample sizes. In general, before analyzing an empirical data set, it is advisable to perform a simulation study to find out whether the results in this article are confirmed for the sample sizes, kind of models and parameter values ranges relevant for the data set at hand.

Worth mentioning here is a recent direction in the area of model selection, that of complexity based criteria (Bozdogan, 1991). In AIC or BIC the complexity of a model is defined as the number k of its estimated parameters. A novel approach proposed by Bozdogan points out that this is too simple a definition of complexity and introduces a new class of criteria called $ICOMP$ (Information Complexity) where the number k is replaced by a suitable measure of the complexity of the model derived from information theoretic principles. $ICOMP$ criteria have been applied to some examples of confirmatory factor analyses (see Bozdogan, 1991), but not in the context of comparing criteria on the basis of frequency of selection of the correct model. For the type of models considered in this study, where the covariance matrix of the errors is diagonal and the number of latent factors equal to three for all analytic models, a suitable $ICOMP$ criterion behaves somewhat like AIC. An interesting direction for a future study would involve simulations containing models with different numbers of factors, possibly a nondiagonal covariance matrix for the measurement errors, and exclusively misspecified (underfitting) analytic models in the selection set.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

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

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332.

Bandalos, D.L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28*(3), 351-374.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P.M., & Bonett, G.B. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Bollen, K.A. (1986). Sample size and Bentler and Bonnet's nonnormed fit index. *Psychometrika, 51*, 375-377.

Bollen, K.A. (1989a). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*, 303-316.

Bollen, K.A. (1989b). *Structural equations with latent variables.* New York: Wiley.

Bollen, K.A. (1990). Overall fit in covariance structure models: two types of sample size effects. *Psychological Bulletin, 107*(2), 256-259.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality.* Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands.

Bozdogan, H. (1987). Model selection and Akaike's in formation criteria (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Bozdogan, H. (1991, June). *A new information theoretic measure of complexity index for model evaluation in general structural equation models with latent variables.* (Paper presented at the Joint Meeting of the Classification Society of North America and Psychometric Society, Rutgers the State University, New Brunswick).

Bozdogan, H., & Haughton, D.M.A. (1995). *Informational complexity criteria for regression models.* (An invited paper for the Special issue on MODEL SELECTION in STATISTICA SINICA, a Journal of the International Chinese Statistical Association).

Browne, M.W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

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. 136-162). Newbury Park, CA: Sage Publications, Inc.

Cudeck, R., & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*, 147-167.

Cudeck, R., & Henly, S.J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*(3), 512-519.

Gerbing, D.W., & Anderson, J.C. (1993). Monte carlo evaluations of good ness-of-fit indices for structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 40-65). Newbury Park, CA: Sage Publications, Inc.

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

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā: The Indian Journal of Statistics, 51*, 45-58.

Haughton, D. (1996). Information criteria for model selection. *Kwantitatieve Methoden* (this issue).

Haughton, D., & Dudley, R.M. (1993, preprint). Information criteria for multiple data sets and restricted parameters.

Haughton, D., Haughton, J., & Izenman, A.J. (1990). Information criteria and harmonic models in time series analysis. *Journal of Statistical Computation and Simulation, 35*, 187-207.

Haughton, D., Oud, J.H.L., & Jansen, R.A.R.G. (1996). *Information and other model selection criteria in LISREL model selection.* (Manuscript submitted for publication).

Hoelter, J.W. (1983). The analysis of covariance structures goodness-of-fit indices. *Sociological Methods & Research, 11*, 325-344.

James, L.R., Mulaik, S.A., & Brett, J.M. (1982). *Causal analysis: assumptions, models, and data.* Beverly Hills, CA: Sage.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In A.S. Goldberger & O.S. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85-112). New York: Seminar Press.

Jöreskog, K.G. (1977). Structural equation models in the social sciences. In P.R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265-287). Amsterdam: North-Holland.

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

Jöreskog, K.G., & Sörbom, D. (1976). *LISREL III: Estimation of linear structural equations systems by maximum likelihood methods.* Chicago: International Educational Services.

Jöreskog, K.G., & Sörbom, D. (1981). *LISREL V: Analysis of linear struc-*

tural relationships by maximum likelihood and least squares methods. Chicago: International Educational Services.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications.* Chicago: SPSS.

Jöreskog, K.G., & Sörbom, D. (1993a). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1993b). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1994). *PRELIS 2: User's reference guide, B: Simulation with PRELIS 2 and LISREL 8.* Chicago: Scientific Software International.

Kass, R. (1989). The geometry of asymptotic inference. *Statistical Science, 4*, 188-234.

Maiti, S.S., & Mukherjee, B.N. (1990). A note on distributional properties of the Jöreskog-Sörbom fit indices. *Psychometrika, 55*, 721-726.

McDonald, R.P., (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*, 97-103.

McDonald, R.P., & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107*(2), 247-255.

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430-445.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 758-765.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*, KTK Scientific Publishers, Tokyo/Kluwer Academic Publishers, USA.

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

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Woodroofe, M. (1982). On model selection and the arc sine laws. *Annals of Statistics, 10*(4), 1182-1194.

Ontvangen: 28 september 1995
Geaccepteerd: 22 maart 1996