Model Selection in Information Systems Research Using Partial Least Squares Based Structural Equation Modeling

Completed Research Paper

Pratyush N. Sharma
Joseph M. Katz Graduate School of Business, University of Pittsburgh
229, Mervis Hall, Pittsburgh, PA 15260
psharma@katz.pitt.edu

Kevin H. Kim
School of Education, University of Pittsburgh
5918, Wesley W. Posvar Hall, Pittsburgh, PA 15260
khkim@pitt.edu

Abstract

Partial Least Squares (PLS) based Structural Equation Modeling (SEM) has become increasingly popular in Management Information Systems (MIS) research to model complex relationships and to make valid inferences from the restricted sample to the larger population. Given the larger goal of creating generalizable theories in MIS research, we argue that the lack of model selection criteria in PLS that penalize model complexity might be causing researchers to select unnecessarily complex but highly fitting models that may not generalize to other samples. We introduce several Information Theoretic (IT) model selection criteria in the PLS context that penalize model complexity but reward high fit, and therefore guide researchers to select a parsimonious and generalizable model. In this Monte Carlo study, we compare their performance to the currently existing PLS indices, in selecting the best model among a set of competing models under various conditions of sample size, effect size and data distribution. Based on our simulation results, we strictly advise against the use of $R^2$ and GoF based measures in PLS for model selection. Instead, we demonstrate that the IT criteria have much superior model selection rates than the currently existing PLS indices. Therefore, we recommend a core set of IT criteria that researchers should regularly employ when selecting models among a competing set of models using PLS based SEM.

Keywords: Partial Least Squares (PLS), Structural Equation Modeling (SEM), Model Selection, Monte Carlo Study
“Everything should be made as simple as possible, but no simpler.” - Albert Einstein

“All Models are wrong, but some are useful!” - George Box

Introduction

Partial Least Squares (PLS) based Structural Equation Modeling (SEM) has become increasingly popular in Management Information Systems (MIS) research to model complex relationships among multiple Latent Variables (LVs), each measured through a number of Manifest Variables. A model is a simplification or approximation of reality and hence does not reflect it in its entirety (Burnham and Anderson, 2002). Models are important because they reflect some (partial) aspect of the reality through the parameters and the relationships in them. Such parameters have relevant, useful interpretations, even when they relate to quantities that are not directly observable, such as the LVs (Burnham and Anderson, 2002). Through such models the MIS community hopes to understand the complex underlying processes and approximate the “true” model that generated the data, and eventually to make valid generalizable inferences from the data. A model should allow us to make valid statistical inferences from the restricted sample data to a real larger population. The model must therefore, not only be rich enough to explain the relations in the data but also be simple enough to understand, explain and use (Claeskens and Hjort, 2010).

PLS has become the technique of choice for many MIS researchers because of its relative simplicity in modeling complicated relationships. In addition, past research has documented several advantages of using PLS over Maximum Likelihood (ML) based SEM techniques. On reviewing the use of PLS in MIS Quarterly over the past 20 year period, Ringle et al. (2012) found that most authors justified the use of PLS due to small sample sizes, non-normal data and the use of formative indicators. Chin (1998) suggested that PLS is suitable for, among other situations, studying phenomena that are relatively new and where the theoretical model is not well formed and/or situations where the model is relatively complex with large number of indicators and LVs. Gefen et al. (2011) also suggested that the superior convergence rates of PLS over ML, mean that PLS can be useful for exploratory research objectives. Because PLS largely avoids inadmissible solutions and factor indeterminacy (Fornell and Bookstein, 1982), it becomes an attractive technique to be used in modeling complex relationships.

In addition, PLS’s prediction orientation means that it aims at maximizing “explained variability” in all endogenous constructs in the structural model. The essential criterion for this assessment is the coefficient of determination (R²) of the endogenous LVs (Henseler et al. 2009). The goodness of fit (GoF) is another R² based index that allows validating a PLS model globally (Tenenhaus et al. 2004). However, the heavy reliance of PLS structural model validation on R² based measures means that when exploring different models, researchers may be tempted to add structural model parameters and relations in an effort to increase R² in one or more LVs and ultimately the GoF (Henseler and Sarstedt, 2012).

In this light, Ringle et al.’s (2012) finding, that PLS models in MIS Quarterly have an average number of about 8 LVs, 27 indicators and 11 structural model relationships, is perhaps not very surprising. By comparison, ML based studies were found to have an average of about 5 LVs and 16 indicators per model (Shah and Goldstein, 2006). Clearly, MIS researchers are using PLS to analyze relatively more complex models than ML. However, this aspect of PLS is as much of a blessing as it has the potential to be a curse, especially when combined with the realization that many MIS researchers justify the use of PLS due to small sample sizes. In our view, the tendency to analyze relatively complex models on restricted sample sizes could be a recipe for disaster. Approximating models must be related to the amount of data and information available; small data sets will appropriately support only simple models with few parameters, while more comprehensive data sets will support, if necessary, more complex models (Burnham and Anderson, 2002). Since both R² and the GoF in their current form do not penalize over-parameterization (Henseler and Sarstedt, 2012), we fear that MIS researchers might be stretching their modeling efforts by analyzing more complicated models than their data would support. We find the absence of indices that...
penalize model complexity in PLS especially troubling since it may be leading MIS researchers against the basic philosophy of Occam’s razor (principle of parsimony) that the scientific discipline rests on.

The goal of exploratory research and model building is to select the “best” approximating model from a set of theoretically justified competing models. However, most SEM studies specify only one model structure and then use the data to either confirm or disconfirm the specific structure (Zheng and Pavlou, 2010). In the context of PLS, selecting a best model from a set of competing models should not only rest on maximizing the explained variance but also on achieving a balance with model parsimony. While the fit ($R^2$) of a PLS model can be increased by increasing the number of parameters and relationships, MIS researchers must guard against this tendency. Instead, an attempt should be made to arrive at a parsimonious model that also fits the data well. A highly complex model can provide a good fit with the data at hand but may not lead to any interpretable true relationship. Myung (2000) showed that model selection solely based on fit to the observed data will result in the choice of an unnecessarily complex model that overfits the data and therefore generalizes poorly.

If MIS researchers give in to the tendency of selecting the model that best fits the data (i.e. which accounts for most variance), they may fall in the trap of selecting an unnecessarily complex model that bears no interpretable relationship with the true underlying process. A complex model due to its additional flexibility may tap spurious patterns in the sample at hand (Myung, 2000). Because such patterns are sample specific, the model may generalize poorly to other samples. Therefore, keeping in mind the larger goal of creating generalizable theories in MIS research, we caution the researchers against the race to increase $R^2$ or GoF. Instead we advocate a strategy similar to Burnham and Anderson (2002), where the researchers develop a manageable set of theoretically justified competing models and then use a set of guiding empirical criteria to select the most parsimonious model that also fits well.

The currently existing set of fit indices in PLS ($R^2$ and GoF) however, have been shown to be woefully inadequate as empirical criteria for model selection. In their recent simulation study, Henseler and Sarstedt (2012) showed that along with $R^2$, the GoF based measures are not suitable for model validation and selection. They showed that neither the GoF nor the relative GoF (GoFrel) were able to separate valid models from invalid ones. The tendency of both $R^2$ and GoF based measures to increase with model complexity (increase in model parameters and relationships) means that these indices will almost always favor complex models over parsimonious ones. Due to these issues, Henseler and Sarstedt (2012) suggested that in order not to be misled in choosing a model with highest GoF, researchers should carefully evaluate path coefficients and their significance in order to decide which paths to leave in the model and which to discard. While carefully evaluating path significance is important, we fear that over reliance on this approach might encourage researchers to cherry-pick significant paths in the model at the cost of theory and its generalizability.

In another recent simulation study, Evermann and Tate (2010) found that PLS model quality criteria displayed a bewildering range of behavior (sic) while evaluating and assessing misspecified models. Due to the weaknesses in the currently existing measures (such as $R^2$ and GoF), Evermann and Tate go so far as to label the use of PLS in exploratory settings without a firm theory base, misleading. Their cautiously pessimistic view runs contrary to the accepted view that PLS is considered more of an exploratory approach rather than a confirmatory one (Vinzi et al. 2010).

With these issues in mind, we introduce and analyze the use of several Information Theoretic (IT) model selection criteria in the PLS context that can help researchers guide their model selection efforts, especially in exploratory settings with an evolving theory base under a set of competing models and hypotheses. All the IT selection criteria described in this study penalize additional model complexity but reward high fit. We specifically argue that selecting a “best” model in the PLS context using these criteria means selecting:

- A model that satisfies all the established PLS model requirements as described by Chin (1998), Marcoulides and Saunders (2006) and Gefen et al. (2011), and
- A model that is closest to the true model (reality) in the Information Theoretic sense and is also parsimonious and thus generalizable.

While the IT selection criteria described here have a solid standing in the regression literature, this is the first study in our knowledge that uses these criteria for model selection in the PLS context. We analyze and compare the performance of the various IT criteria and the currently existing PLS indices, in selecting...
the best model under various conditions of sample size, effect size and data distribution, using a Monte Carlo study. In what follows, we give a brief background of the various IT criteria, describe our research methodology, analyze the results, describe limitations of the current study and conclude with implications for researchers.

**Background**

The research on developing methods to select the best model among a set of competing models has a distinguished history in the regression literature. The simplest model selection criterion that could be used is the $R^2$ which is calculated as:

$$R^2_k = 1 - \frac{SSerror(k)}{SS\text{total}}$$

where, $SSerror(k)$ is the sum of squares error for the $k^{th}$ model in a set of models and $SS\text{total}$ is total sum of squares. Given that $R^2$ will generally increase whenever extra predictors are added to the model and hence will select a more complex model, regression researchers have widely used the Adjusted $R^2$, which attempts to correct for model complexity. It is given by:

$$\text{Adj. } R^2_k = 1 - \left(1 - \frac{n - 1}{n - p_k}\right) \left(\frac{SSerror(k)}{SS\text{total}}\right)$$

where, $p_k$ is the number of predictors in the $k^{th}$ model, plus 1. Unlike $R^2$, the Adj.$R^2$ does not necessarily increase with the number of predictors. In the late 60s and the early 70s however, other model selection criteria began to appear in the literature that penalized model complexity in the interest of the principle of parsimony. McQuarrie and Tsai (1998) categorize the development of these criteria in two parallel streams of work that differed in the underlying assumptions they made regarding the knowledge of the true model that generated the data.

The first stream of selection criteria assumed that the true model (reality) is of infinite dimension and is essentially unknown. These selection criteria included Akaike’s (1969) Final Prediction Error (FPE), Mallows Cp (Mallows, 1973), Akaike’s (1973) Information Criterion (AIC), Sugiura’s (1978) corrected AIC (AICc) and McQuarrie and Tsai’s (1998) unbiased AIC (AICu). Due to the realistic assumption that the true model is unknown, these criteria strive to select a model that is closest to the unknown true model in the sense of some well defined distance measure. The earliest selection criteria, FPE and Cp both relied on the $L_2$ norm as the basis of measuring the distance. The $L_2$ norm is defined as:

$$\Delta L_2(M_t, M_a) = \|\mu_{M_t} - \mu_{M_a}\|^2$$

where, $M_t$ is the true model and $\mu_{M_t}$ is its mean, and $M_a$ is the candidate model with a mean $\mu_{M_a}$ respectively. On the other hand, the AIC based criteria were based on the Information Theoretic notion of the Kullback-Liebler (KL) discrepancy to measure the distance, which is defined as:

$$\Delta KL(M_t, M_a) = E_{F_t}\left[\log \frac{f_t(x)}{f_a(x)}\right]$$

where, $M_t$ is the true model with density $f_t(x)$ and distribution $F_t$ and $M_a$ is the candidate model. The main advantage of these criteria was that they could be used to measure the relative distances of competing models from the unknown true model, even when the absolute distances to the true model could not be known. This allowed researchers to compare the relative distances and select a model that was closest to the unknown true model as compared to other competing models. This involved selecting the model with the smallest value on these criteria. In large samples, a criterion that selects a model with minimum mean squared error is considered asymptotically efficient (Shibata, 1980). All the distance based criteria discussed here, FPE, Cp, AIC, AICc and AICu are asymptotically efficient (McQuarrie and Tsai, 1998). Table 1 presents the details of these criteria.
Table 1. Distance Based Efficient Criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Formula</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Prediction Error (FPE)</td>
<td>$\left( \frac{SS_{error}}{n - p_k} \right) \times \left( \frac{1 + p_k}{n} \right)$</td>
<td>Selects the best model by minimizing the final prediction error.</td>
</tr>
<tr>
<td>Mallow’s Cp</td>
<td>$\left( \frac{SS_{error}}{MS_{error}} \right) - (n - 2p_k)$</td>
<td>Based on mean square error (MSE) of prediction; $MS_{error}$ is MSE from the saturated (full) model.</td>
</tr>
<tr>
<td>Akaike Information Criterion (AIC)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n} \right) + \frac{2p_k}{n} \right]$</td>
<td>Estimates the relative expected KL distance to the unknown true model.</td>
</tr>
<tr>
<td>Unbiased AIC (AICu)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n - p_k} \right) + \frac{2p_k}{n} \right]$</td>
<td>Uses the unbiased estimate for population MSE, hence differs from AIC in small samples.</td>
</tr>
<tr>
<td>Corrected AIC (AICc)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n} \right) + \frac{n + p_k}{n - p_k - 2} \right]$</td>
<td>Corrects AIC’s tendency to overfit (select a complicated model) under small samples.</td>
</tr>
</tbody>
</table>

The second stream of model selection criteria came from researchers who assumed that the true model is finite, fully known and included in the set of competing models. The goal of the selection criteria based on this assumption is to choose the correct model among the competing models, such as by maximizing the posterior model probability (BIC). Therefore, these criteria do not rely on any measure of distance. Any model selection criterion that always correctly selects the true model is considered to be consistent (McQuarrie and Tsai, 1998). Examples of such criteria include Akaike (1978) and Schwarz’ (1978) Bayesian Information Criterion (BIC), Geweke and Meese’s (1981) criterion (GM), Hannan and Quinn’s (1979) criterion (HQ) and McQuarrie and Tsai’s (1998) corrected HQ criterion (HQc). BIC, GM, HQ and HQc are all asymptotically consistent (McQuarrie and Tsai, 1998). Generally researchers select a model with the least value on these criteria. Table 2 presents the details of these criteria.

In all the IT criteria mentioned in Tables 1 and 2, the first term can be interpreted as a measure of lack of model fit while the second term can be interpreted as the penalty for increasing model complexity (Burnham and Anderson, 2002). Therefore, all of these criteria try to achieve a trade-off between model fit and model parsimony.

Table 2. Consistent Criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Formula</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Information Criterion (BIC)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n} \right) + \frac{p_k \log(n)}{n} \right]$</td>
<td>Derived using Bayesian argument; adjusts AIC for model complexity by using a stronger penalty for overfitting.</td>
</tr>
<tr>
<td>Geweke-Meese Criterion (GM)</td>
<td>$\left( \frac{SS_{error}}{MS_{error}} \right) + p_k \log(n)$</td>
<td>Adjusts Mallow’s Cp for model complexity by using a stronger penalty for overfitting.</td>
</tr>
<tr>
<td>Hannan-Quinn Criterion (HQ)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n} \right) + \frac{2p_k \log \log(n)}{n} \right]$</td>
<td>Corrects small sample performance of BIC by using a stronger penalty term.</td>
</tr>
<tr>
<td>Corrected HQ Criterion (HQc)</td>
<td>$n \left[ \log \left( \frac{SS_{error}}{n} \right) + \frac{2p_k \log \log(n)}{n - p_k - 2} \right]$</td>
<td>Corrects small sample performance of HQ and adjusts for model complexity.</td>
</tr>
</tbody>
</table>

When comparing the efficient and consistent criteria, researchers have reached little agreement on which of these criteria are better and the choice remains highly subjective and researcher specific (McQuarrie and Tsai, 1998). Research efforts to combine both these properties in a single criterion have so far not
been entirely successful either. For example, Yang (2005) sought to combine the efficiency of AIC with the consistency of BIC. Instead, he found that there exists an inherent trade-off such that if one pursued one aspect, one must sacrifice the other.

Further, since efficiency and consistency are both asymptotic properties, the behavior of these criteria under the various conditions that MIS researchers commonly work under needs to be explored. Specifically, the behavior of these criteria for PLS model selection under various conditions of sample size, effect size and distributional conditions are unknown. For this paper, we conducted a Monte Carlo study to explore these issues further and to create a helpful guideline for MIS researchers using PLS based SEM for model selection. In the next section, we elaborate on our methodology for the simulations.

Methodology

We compared the model selection performance of PLS based fit indices ($R^2$, Adj. $R^2$, GoF and Stone-Geisser’s $Q^2$) and the various IT selection criteria as described in Tables 1 and 2. In PLS based SEM, the structural (inner) model represents the theoretically hypothesized relationships among the LVs. We created a set of 7 competing structural models, one of which was the true model that generated the data. These 7 models could be thought of as representing sets of competing hypotheses that a researcher may want to explore. All structural models in our study had 4 LVs, two of which were exogenous ($\xi_1$ and $\xi_2$) and the other were endogenous ($\eta_1$ and $\eta_2$). Since the focus of our study is the selection of the true structural (inner) model, we fixed the measurement (outer) model to be the same across all the 7 models. Each LV had 3 reflective indicators with fixed loadings of 0.6. The focus of our investigation is the amount of variance explained in the exogenous variable $\eta_2$. The seventh model was a fully saturated model with all possible relationships predicting $\eta_2$, as shown in Figure 1.

All the other models were a subset of this saturated model. We created these 7 models through all possible combinations of paths predicting the exogenous variable $\eta_2$ (A, B and C in Figure 1). These competing models are described in Table 3. While this approach of creating competing models through “all possible subsets” is reasonable for a simulation study, we strongly advocate against this approach while testing actual competing hypotheses. Researchers should create a manageable subset of competing models guided by existing theory and not through a brute-force approach that models all possible subsets. A model that does not make theoretical sense should not be included in this set. The saturated model is the most parameterized model possible and serves as the basis for assessing overall fit. If the saturated model fits well, then a parsimonious model that is a subset will also generally fit well (Burnham and Anderson, 2002).
Table 3. Competing Models

<table>
<thead>
<tr>
<th>Paths predicting $\eta^2$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>A</td>
</tr>
<tr>
<td>Model 2</td>
<td>B</td>
</tr>
<tr>
<td>Model 3</td>
<td>C</td>
</tr>
<tr>
<td>Model 4</td>
<td>A, B</td>
</tr>
<tr>
<td>Model 5</td>
<td>A, C</td>
</tr>
<tr>
<td>Model 6</td>
<td>B, C</td>
</tr>
<tr>
<td>Model 7</td>
<td>A, B, C</td>
</tr>
</tbody>
</table>

Data were generated using Fleishman’s method (Fleishman, 1978; Vale and Maurelli, 1983) for the underlying true model under 6 conditions of sample size (50, 100, 150, 200, 250 and 500), 5 conditions of varying effect size on path A (0, 0.1, 0.2, 0.3 and 0.5) and 4 distributions (normal, chi-squared distributed with $df = 3$, $t$ distributed with $df = 5$, and uniform). These distributions were chosen to reflect normal, positively skewed, heavy tailed and uniform distributions, respectively. All simulations were run on the R computing environment (R Development Core Team, 2011) using the sempls package (Monecke, 2012). Two hundred dataset replications were performed for each of the 120 conditions. The dependent variable of interest was a binary variable that assumed the value 1 if the model selection criterion selected the true model; otherwise it assumed the value 0. In the case of PLS based fit indices, a true model was selected if the fit index achieved the highest value among the competing models. In the case of IT selection criteria, a true model was selected if the criterion achieved the lowest value among the competing models.

Analysis and Results

Due to the large number of experimental conditions ($6 \times 5 \times 4 = 120$), providing the individual results across all conditions would be cumbersome. Instead, we present the descriptive results of the mean true model identification rates of each criterion across all the 120 conditions (means of cell means). Table 4 presents these descriptive results. The worst case scenario (Min.) represents one (of the 120) experimental conditions where the criterion performed the worst across the 200 samples for that condition; while the best case scenario (Max.) represents one (of the 120) experimental conditions where the criterion performed the best across the 200 samples for that condition.

As can be seen in Table 4, in terms of the PLS based indices, $R^2$ and GoF performed very poorly as compared to Adj. $R^2$ and $Q^2$ in selecting the true model. In the best case scenario, $R^2$ and GoF were able to identify the true model only 10% and 23% of times respectively. In the worst case scenario, they were not able to identify the true model even once. Their average identification rates (Mean), across all the 120 experimental conditions, were also very low (3% and 12%). This is perhaps not surprising since both would favor the saturated model over the parsimonious true model. $Q^2$ performed the best overall (average identification rate 47%), followed by Adj. $R^2$ (average identification rate 40%).

In terms of the IT criteria, we saw a far wider range of percentage correct identifications when comparing the best and worst case scenario of each criterion. This can be seen in the high standard deviations of these criteria. For example, in the best case scenario both GM and BIC were able to identify the true model almost all the time (93%), while under the worst case scenario they were able to identify the true model only 4% of the times. This suggests that the experimental conditions affected these IT criteria more severely than others. Overall, HQ and HQc performed the best (average identification rate 48%), closely followed by $Q^2$, FPE, Cp and the AIC based criteria (47%). Clearly, among the PLS based indices only $Q^2$ was able to compete with the higher performance rates of IT criteria in terms of model selection. Therefore, we strongly recommend against the use of $R^2$, Adj. $R^2$ and GoF for model selection purposes in PLS based studies.
In order to analyze the pattern of effect of experimental conditions (sample size, effect size and distribution types) on the performance of the model selection criteria, we created 3 separate two-way marginal means tables with the selection criteria as one dimension. Table 5 presents the results for sample size and the various selection criteria.

Table 5. Selection Criterion vs. Sample Size (Percentage correctly identified)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>R²</th>
<th>Adj. R²</th>
<th>GoF</th>
<th>Q²</th>
<th>FPE</th>
<th>Cp</th>
<th>GM</th>
<th>AIC</th>
<th>AICu</th>
<th>AICc</th>
<th>BIC</th>
<th>HQ</th>
<th>HQc</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.05</td>
<td>.24</td>
<td>.12</td>
<td>.23</td>
<td>.22</td>
<td>.22</td>
<td>.16</td>
<td>.22</td>
<td>.18</td>
<td>.21</td>
<td>.15</td>
<td>.19</td>
<td>.18</td>
</tr>
<tr>
<td>100</td>
<td>.04</td>
<td>.37</td>
<td>.13</td>
<td>.40</td>
<td>.39</td>
<td>.39</td>
<td>.29</td>
<td>.39</td>
<td>.35</td>
<td>.39</td>
<td>.28</td>
<td>.35</td>
<td>.34</td>
</tr>
<tr>
<td>150</td>
<td>.03</td>
<td>.41</td>
<td>.13</td>
<td>.48</td>
<td>.47</td>
<td>.47</td>
<td>.39</td>
<td>.47</td>
<td>.46</td>
<td>.47</td>
<td>.38</td>
<td>.45</td>
<td>.45</td>
</tr>
<tr>
<td>200</td>
<td>.04</td>
<td>.45</td>
<td>.12</td>
<td>.53</td>
<td>.54</td>
<td>.54</td>
<td>.50</td>
<td>.54</td>
<td>.56</td>
<td>.55</td>
<td>.49</td>
<td>.56</td>
<td>.56</td>
</tr>
<tr>
<td>250</td>
<td>.03</td>
<td>.49</td>
<td>.13</td>
<td>.58</td>
<td>.59</td>
<td>.59</td>
<td>.57</td>
<td>.59</td>
<td>.62</td>
<td>.59</td>
<td>.56</td>
<td>.62</td>
<td>.62</td>
</tr>
<tr>
<td>500</td>
<td>.02</td>
<td>.43</td>
<td>.09</td>
<td>.59</td>
<td>.60</td>
<td>.60</td>
<td>.75</td>
<td>.60</td>
<td>.68</td>
<td>.60</td>
<td>.75</td>
<td>.71</td>
<td>.72</td>
</tr>
<tr>
<td>Overall</td>
<td>.03</td>
<td>.40</td>
<td>.12</td>
<td>.47</td>
<td>.47</td>
<td>.47</td>
<td>.44</td>
<td>.47</td>
<td>.48</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
</tbody>
</table>

We can see that sample size affected the IT criteria more than it did the PLS based indices. In fact, R² and GoF were hardly affected by an increase in sample size and their performance stayed very poor throughout, while Q² again emerged as the best among the PLS based indices. In terms of the IT based criteria, GM and BIC achieved correct identification rates of 75% at sample size 500. These results suggest that at smaller sample sizes (less than 200) researchers can hardly place much confidence in any of these measures for model selection. Only at sample sizes 200 and above did we see that the IT selection criteria begin to select the true model with some regularity. Therefore, we caution MIS researchers against justifying the use of PLS in exploratory studies involving competing models with smaller sample sizes.

Next, we turned to analyzing the effect of varying effect size (path A in our model) and the various selection criteria. Table 6 presents the results. Note that at effect size 0 (path A loading = 0), the selection
criteria should select the true model 3; while for all other effect sizes the true model to be selected is the model 5.

### Table 6. Selection Criterion vs. Effect Size (Percentage correctly identified)

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>( R^2 )</th>
<th>Adj. ( R^2 )</th>
<th>GoF</th>
<th>( Q^2 )</th>
<th>FPE</th>
<th>( Cp )</th>
<th>GM</th>
<th>AIC</th>
<th>AICu</th>
<th>AICc</th>
<th>BIC</th>
<th>HQ</th>
<th>HQc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>.01</td>
<td>.28</td>
<td>.01</td>
<td>.53</td>
<td>.46</td>
<td>.46</td>
<td>.65</td>
<td>.46</td>
<td>.56</td>
<td>.48</td>
<td>.66</td>
<td>.58</td>
<td>.59</td>
</tr>
<tr>
<td>0.1</td>
<td>.05</td>
<td>.33</td>
<td>.12</td>
<td>.28</td>
<td>.29</td>
<td>.29</td>
<td>.11</td>
<td>.29</td>
<td>.22</td>
<td>.28</td>
<td>.10</td>
<td>.21</td>
<td>.20</td>
</tr>
<tr>
<td>0.2</td>
<td>.04</td>
<td>.43</td>
<td>.14</td>
<td>.44</td>
<td>.46</td>
<td>.46</td>
<td>.29</td>
<td>.46</td>
<td>.41</td>
<td>.45</td>
<td>.29</td>
<td>.41</td>
<td>.39</td>
</tr>
<tr>
<td>0.3</td>
<td>.04</td>
<td>.47</td>
<td>.15</td>
<td>.53</td>
<td>.54</td>
<td>.54</td>
<td>.48</td>
<td>.54</td>
<td>.54</td>
<td>.54</td>
<td>.47</td>
<td>.55</td>
<td>.54</td>
</tr>
<tr>
<td>0.5</td>
<td>.03</td>
<td>.47</td>
<td>.17</td>
<td>.58</td>
<td>.59</td>
<td>.59</td>
<td>.66</td>
<td>.59</td>
<td>.64</td>
<td>.60</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>Overall</td>
<td>.03</td>
<td>.40</td>
<td>.12</td>
<td>.47</td>
<td>.47</td>
<td>.47</td>
<td>.44</td>
<td>.47</td>
<td>.48</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
</tbody>
</table>

For effect size equal to 0, \( Q^2 \) and the IT selection criteria did reasonably well in selecting the true model (model 3), while \( R^2 \), Adj. \( R^2 \) and GoF performed poorly. For effect sizes larger than 0, we again found that \( R^2 \) and GoF performed very poorly as compared to Adj. \( R^2 \), \( Q^2 \) and the IT criteria while selecting the true model (model 5). With an increase in the effect size on path A, the ability of \( R^2 \) and GoF to select the true model hardly increased. While the performance rate of Adj. \( R^2 \) quickly tapered off around effect size 0.3, the performance of \( Q^2 \) stayed competitive as compared to the high performing IT criteria. GM, BIC, HQ, and HQc emerged as the most successful selection criteria at effect sizes of 0.3 or greater. Overall, an increase in the effect size was associated with an increase in the selection criteria’s abilities to select the true model. Next, we analyzed the effect of different data distributions and the various selection criteria. Table 7 presents the results.

### Table 7. Selection Criterion vs. Dist. Types (Percentage correctly identified)

<table>
<thead>
<tr>
<th>Dist. Type</th>
<th>( R^2 )</th>
<th>Adj. ( R^2 )</th>
<th>GoF</th>
<th>( Q^2 )</th>
<th>FPE</th>
<th>( Cp )</th>
<th>GM</th>
<th>AIC</th>
<th>AICu</th>
<th>AICc</th>
<th>BIC</th>
<th>HQ</th>
<th>HQc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>.04</td>
<td>.39</td>
<td>.12</td>
<td>.48</td>
<td>.47</td>
<td>.47</td>
<td>.45</td>
<td>.47</td>
<td>.56</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
<tr>
<td>Chi-Sq</td>
<td>.03</td>
<td>.40</td>
<td>.12</td>
<td>.47</td>
<td>.47</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.22</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
<tr>
<td>t-dist</td>
<td>.04</td>
<td>.40</td>
<td>.12</td>
<td>.47</td>
<td>.46</td>
<td>.46</td>
<td>.43</td>
<td>.47</td>
<td>.41</td>
<td>.46</td>
<td>.43</td>
<td>.48</td>
<td>.47</td>
</tr>
<tr>
<td>Uniform</td>
<td>.03</td>
<td>.40</td>
<td>.11</td>
<td>.46</td>
<td>.47</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.54</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
<tr>
<td>Overall</td>
<td>.03</td>
<td>.40</td>
<td>.12</td>
<td>.47</td>
<td>.47</td>
<td>.47</td>
<td>.44</td>
<td>.47</td>
<td>.43</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.48</td>
</tr>
</tbody>
</table>

In our simulation we had tried to capture a wide range of data distributions that MIS researchers are likely to face in their studies. These distributions ranged from normal, positively skewed, heavy tailed to uniform distributions. The results of the Table 7 strongly suggest that the performance of the selection criteria remained stable over changes in data distributions. Since PLS is generally considered robust against non-normal distributions, these results are in line with the established PLS literature. We therefore see this as an advantage when using the IT selection criteria for exploratory PLS studies where meeting rigid assumptions about data distribution is hardly feasible.

Finally, to confirm if there were significant interaction effects among the selection criteria and the various experimental conditions (sample size, effect size and data distributions), we performed a repeated measures logistic regression analysis using the Generalized Estimating Equation technique. This regression was performed on the true model selection outcome, as predicted by the selection criteria, sample size, effect size and distribution type. The results are shown in Table 8.
The results in Table 8 suggest significant interaction effects between the selection criteria and sample size and effect size respectively. The interaction effect of selection criteria and distribution types was not significant. The regression results confirm our analysis of the behavior of the selection criteria above.

**Limitations and Future Work**

Like all research our study is limited in several ways. First, all our variables were generated on a continuous scale, while in practice researchers most often work with categorical and nominal data. Second, we have focused our efforts only on analyzing the structural model by keeping the measurement model constant throughout. As such, our results are strictly generalizable to situations where the researchers have a similar measurement model. We leave it for future researchers to explore the effects of measurement model manipulations and how they may interact with structural model manipulations while selecting a best approximating model. In addition, our choice of setting the factor loadings to 0.6 seemed appropriate and realistic for an exploratory context, where researchers typically deal with less developed and possibly inferior measures. As the measures get more accurate and the factor loadings increase, noise is reduced and more power is achieved. Therefore, we speculate that with an increase in the factor loading values, the IT measures would perform with even greater accuracy. However, this contention needs to be verified through future studies.

Another potential avenue for research could be to assess the performance of the IT criteria under different conditions of the indicators-to-factors ratio, which has been shown to have an impact in the covariance based SEM (CB-SEM) context (Homburg, 1991). It would be worthwhile to ascertain if the same holds true in the PLS based SEM context. While CB-SEM and PLS-SEM indeed have many similarities, there are major differences among these techniques, especially in their intended focus. CB-SEM focuses on estimating a set of model parameters such that the theoretical covariance matrix is as close as possible to the sample covariance matrix (Reinartz et al., 2009). On the other hand, PLS focuses on maximizing the explained variance ($R^2$) for all the endogenous constructs. Since the IT criteria also rely on $R^2$ in their measurement, we consider PLS-SEM as a more suitable context for their use. Even so, we see the need to compare the performance of these IT criteria, in terms of the correct identification of the “best” model, to the performance of various fit indices commonly used in CB-SEM.

Furthermore, our models are all simple mediation models without any moderation effects. As such our results are strictly generalizable only to similar models. Although we speculate that our results should be generalizable to more complex models, we regard this as an open issue to be resolved by future studies.

Finally, the equations to calculate the IT criteria as presented in tables 1 and 2, assume normal distribution of the variables. In the strictest sense, if the variables follow a non-normal distribution then the IT criteria should be adjusted and calculated based on the maximum value of the likelihood function. However, our results showed that the IT criteria as presented in tables 1 and 2 indeed produced robust and valid results. We leave it for future research to assess whether there are significant gains to be

---

1 We thank the anonymous reviewers for pointing out these important issues.
achieved by using the adjusted versions of the IT criteria over the criteria presented in tables 1 and 2, and under which conditions.

Implications

In this study we showed that the PLS based indices, namely $R^2$, Adj. $R^2$ and GoF are unsuitable for model selection purposes. This is especially true when researchers need to analyze several potential structural models in exploratory settings. We introduced the use of several Information Theoretic (IT) model selection criteria in PLS based SEM. Utilizing these criteria, researchers can explore competing hypotheses with a larger empirical tool set than is currently available in PLS. Based on our simulation results (Tables 4-7), we propose the following guidelines for MIS researchers regarding the use of model selection criteria in PLS based SEM:

- **Strictly not recommended** – We strictly do not recommend using either $R^2$ or GoF for model selection purposes. The performances of these criteria were very poor while selecting the true model among the set of competing models under all conditions. This could be due to their tendency to overfit and select a more complicated model than a parsimonious true model.

- **Somewhat recommended** – In terms of the PLS based indices, we found that $Q^2$ was the best performing model selection criterion, followed by Adjusted $R^2$. Even though Adjusted $R^2$ attempted to correct for the increase in model complexity, it was still unable to compete with $Q^2$ and the IT criteria. Therefore, we would recommend researchers to prefer $Q^2$ over Adjusted $R^2$, if they do decide to use a PLS based index for model selection. In terms of the IT criteria, we found that FPE, Cp, AIC and AICc performed reasonably well under some conditions and among the PLS based indices only $Q^2$ was able to give a comparable performance.

- **Highly recommended** – Finally, we recommend the core set of high performing selection criteria that MIS researchers should regularly employ in their model selection endeavors. This set includes the GM, AICu, BIC, HQ and HQc criteria. All the criteria in this set have comparable performances and MIS researchers can use these criteria to improve their confidence level while selecting a best model from a set of competing models. We believe that with the use of this additional tool set, MIS researchers will be able to employ PLS with additional confidence under exploratory settings, where the technique is generally considered to be superior to others.

Our results regarding the above mentioned recommended set are consistent with the multiple regression literature (McQuarrie and Tsai 1998). Since PLS also rests on the OLS framework, our findings strongly suggest that these criteria can be fruitfully employed in this context. As per our knowledge, however, none of the recommended IT criteria are currently provided by any PLS software in a model selection context, although SmartPLS (Ringle et al. 2005) does provide AIC and BIC in the context of sample segmentation using FIMIX PLS. Therefore, we would highly recommend that provisions be made so that these criteria are made available for researchers to be used in the model selection context.

There are several advantages of using the aforementioned recommended IT criteria. First, all the IT criteria discussed here penalize model complexity which the current PLS indices are not designed to do. This might help MIS researchers to select potentially more parsimonious models that not only fit the data well, but also generalize to other samples. Second, researchers can use the IT criteria to compare even non-nested models. Since the IT criteria are not based on the traditional hypothesis testing paradigm, they are not bound by the nesting requirement. Third, the order of testing the competing models is irrelevant to proper interpretation, unlike the hypothesis testing approaches (such as, forward addition, backward elimination and stepwise), that may be affected by the order (Burnham and Anderson, 2002).

Next, we extend some cautious remarks to researchers when using the IT criteria for model selection. First, the IT criteria values can only be used to compare models on the same data set. Researchers cannot hope to compare and interpret models across samples using these IT criteria. Second, all IT criteria assume that some endogenous LV (outcome) is the subject of interest hence it only makes sense to compare models predicting the same LV (Burnham and Anderson, 2002). Third, these criteria were derived using Information Theory which represents a different analysis paradigm than the traditional hypothesis testing paradigm. Therefore, they cannot be used for hypothesis testing and thus it would be incorrect to claim that one model is “significantly” better than the other (Burnham and Anderson, 2002). Fourth, researchers should generate a manageable set of competing models by taking guidance from the
existing theory. Clearly, if a model does not make theoretical sense it should not be included in the set of competing models. Theory will also be helpful in separating models that are too close in the IT sense. Fifth, we have shown that the performance of these criteria is adversely affected at smaller sample sizes (less than 200) and at smaller effect sizes (less than 0.3). A small sample contains relatively little information and hence we caution researchers while using these criteria with restricted sample sizes.

Finally, the use of IT criteria allowed us to improve the identification rates of the best model from the “base rate” of about 14% (correctly identifying a best model out of 7 competing models by chance) to about 60% with normal sample sizes (200-250). Tellingly, the identification rates provided by R^2 and GoF were even worse than the base rate. Therefore, while the use of IT criteria does represent a significant improvement over the PLS-SEM measures, these criteria are still not entirely fail-proof. As a consequence, a potential dilemma that a researcher is likely to face could be the conflicting recommendations provided by the IT criteria. In such a case, we would consider it wise to follow the recommendation of the majority of the 5 highly recommended criteria (GM, AICu, BIC, HQ and HQc). In case the researcher finds that none of the competing models are being favored by the majority of the core criteria, it would be prudent to recognize that the researcher has multiple competing models that work equally well. In such a circumstance, the role of theory should be considered paramount in favoring one of the multiple competing models over others, until such time as additional data samples could be collected for further testing and validation.

**Conclusion**

In this study, we argued that the current set of PLS based fit indices such as R^2 and GoF may erroneously lead MIS researchers to select over-parameterized models. Given the larger goal of creating generalizable theories in MIS research, we argued that the lack of PLS model selection criteria that penalize model complexity might be causing researchers to select unnecessarily complex but highly fitting models that may not generalize to other samples. Therefore, we strictly do not recommend the use of R^2 and GoF for model selection purposes. Should the researchers decide to use a PLS based index, we recommend that the best choice would be Q^2 which should be preferred over Adjusted R^2. We introduced various Information Theoretic (IT) criteria that penalize over-parameterization of models and try to achieve a trade-off between fit and model parsimony. We showed that these criteria performed much better while selecting the true model underlying the data, than the currently available fit indices in PLS. Based on our simulation results, we highly recommend a core set of high performing IT criteria that MIS researchers should regularly employ when comparing models under exploratory settings. This core set includes the GM, AICu, BIC, HQ and HQc criteria. Finally, we showed that the true model selection performances of these criteria are adversely affected at smaller sample sizes (less than 200) and small effect sizes (less than 0.3), but not by the non-normal distribution of the data. Therefore, with a proper theoretical base and a strong study design, these criteria should allow researchers to analyze competing models and select the best model among them. While certainly not a panacea, we believe that these IT criteria can definitely provide the much needed empirical guidance to MIS researchers while using PLS to make more informed model selection decisions under exploratory settings.

**References**


