Predictive model selection in partial least squares path modeling

Research-in-Progress

Pratyush Nidhi Sharma
University of Delaware
pnsharma@udel.edu
Galit Shmueli
National Tsing Hua University
galit.shmueli@iss.nthu.edu.tw
Marko Sarstedt
Otto-von-Guericke-University Magdeburg
marko.sarstedt@ovgu.de
Kevin H. Kim (†)
University of Pittsburgh

Abstract

Predictive model selection metrics are used to select models with the highest out-of-sample predictive power among a set of models. R^2 and related metrics, which are heavily used in partial least squares path modeling, are often mistaken as predictive metrics. We introduce information theoretic model selection criteria that are designed for out-of-sample prediction and which do not require creating a holdout sample. Using a Monte Carlo study, we compare the performance of frequently used model evaluation criteria and information theoretic criteria in selecting the best predictive model under various conditions of sample size, effect size, loading patterns, and data distribution.

Keywords: *Partial Least Squares Path Modeling (PLS-PM), Structural Equation Modeling (SEM), Out-of-Sample Prediction, Model Selection, Monte Carlo Study*.

1. Introduction

As researchers our quest is to describe natural processes that interest us with fidelity and economy. We strive to create models that parsimoniously describe reality in the hope that they are generalizable across contexts. Several notions of generalizations exist; the first is *statistical generalization*, where the model estimated from the sample generalizes to the population from which the sample was drawn. In such a case, fitting the same model to a different sample from the same population should yield a similar model. A second type is *scientific generalization*, where the model estimated from the sample generalizes to other populations (e.g., to other contexts). A third type, which is the goal of this paper, is *predictive generalization* where the model estimated from the sample provides sufficiently accurate predictions for new records from that population (out-of-sample prediction). Using a model to generate out-of-sample predictions for new observations is both practically useful as well as essential for scientific model development. Predictive power is useful for assessing the relevance of models, for comparing competing theories, for developing new measures, and more (Shmueli and Koppius, 2011).

In recent years, partial least squares path modeling (PLS-PM) has become increasingly popular in various disciplines to model complex relationships among multiple latent variables, each measured through a number of manifest variables (e.g., Hair et al., 2012a, b; Lee et al., 2011; Ringle et al., 2012). As a composite-based method, PLS-PM has an advantage over factor-

based structural equation modeling (SEM) methods because it yields determinate predictions. Thus, PLS-PM trades optimality for flexibility and the ability to predict (Becker et al., 2013). The ability to predict is one of the most common arguments for using PLS-PM over factor-based SEM, even though most researchers never use any predictive metrics (such as Q^2) or hold-out samples to measure the actual out-of-sample prediction abilities of their models (Ringle et al., 2012). In contrast, the use of R^2 and related measures, which are often mistaken as predictive rather than measures of in-sample explanatory power (Shmueli and Koppius, 2011), is highly common. More precisely, of all the 532 models analyzed in Hair et al.'s (2012a, b) and Ringle et al.'s (2012) reviews of PLS-PM use, 470 models (88.35%) report the R^2, 54 models (10.15%) the Q^2, and 16 models (3.01%) the goodness-of-fit index (GoF), another R^2-based model evaluation measure (Tenenhaus et al., 2004). This almost exclusive focus on a model's explanatory power or in-sample prediction power is problematic as measures such as R^2 or GoF improve with the model's complexity. As a consequence, these indices will almost always favor complex models over simpler ones. In this light, Ringle et al.'s (2012) finding, that PLS models in *MIS Quarterly* are much more complex (e.g., in terms of the number of structural model relationships) compared to those used in factor-based SEM studies in related disciplines is not surprising. This general trend toward more complex PLS models is not restricted to the management information systems field. In marketing and strategic management literatures, Hair et al. (2012a, b) reported similar findings. More importantly, both author groups saw a significant increase in PLS model complexity for papers published in top journals since 2000. Thus, researchers are not only using PLS-PM to analyze more complex models than factor-based SEM, but they are also increasingly testing more complex models than ever before. While the focus on maximizing explanatory power and in-sample prediction using complex models is a worthy goal, there is a real risk that an over-reliance on corresponding measures might tempt researchers to *overfit* their models when their goal is in fact out-of-sample prediction and replicability.

It is a well-confirmed fact among statisticians and applied mathematicians that more complex models often predict poorly out of sample (e.g. Forster and Sober, 1994; Hitchcock and Sober, 2004). A complex model, due to its additional flexibility, might tap spurious patterns in a sample (Myung, 2000). Because such patterns are sample-specific, an overly complex (i.e., overfitted) model will predict poorly and may not be generalizable or replicable by other researchers. In contrast, models with fewer parameters stand a better chance of having higher predictive power and being scientifically replicable (Bentler and Mooijaart, 1989). Thus, researchers using PLS-PM should be aware of the trade-off that exists between model complexity and predictive accuracy. Akaike (1973) showed that this trade-off is achievable, and that an unbiased estimate of a model's out-of-sample predictive accuracy can be obtained by taking into account the fit to the data as well as the model's simplicity. Thus, parsimony plays a crucial role in defining predictive accuracy as a goal in model selection (Hitchcock and Sober, 2004). Therefore, researchers interested in models with predictive power should develop a manageable set of theoretically motivated competing models and then use a set of out-of-sample prediction criteria to select a model that offers the best compromise with model fit and parsimony (Burnham and Anderson, 2002).

Recent research has started to systematically explore PLS-PM's out-of-sample predictive capabilities. For example, Becker et al. (2013) examined the predictive ability of PLS-PM with models including formative constructs, using a modified version of the R^2—which involves a comparison of sample and population composite scores—as a criterion. Evermann and Tate (2014) recently extended this study by comparing out-of-sample prediction of PLS-PM with a

range of different methods, including CBSEM. While both studies make valuable contributions to the literature on PLS-PM, their focus is on researching the method's predictive capabilities. Correspondingly, both author groups rely on a limited set of out-of-sample prediction criteria, which do not directly penalize model complexity and require the construction of a holdout sample for model comparison and selection. Our aim in this paper is to provide researchers with model selection criteria that are tuned to out-of-sample prediction yet do not require a holdout set. Selecting the model with the highest out-of-sample prediction power among a set of potential models is especially useful in exploratory analysis. PLS-PM typically involves some level of exploration. The technique's originator characterized the process of PLS modeling as follows (Wold 1980, pp. 70): "The arrow scheme is usually tentative since the model construction is an evolutionary process. The empirical content of the model is extracted from the data, and the model is improved by interactions through the estimation procedure between the model and the data and the reactions of the researcher."

With this issue in mind, our paper introduces predictive information theoretic model selection criteria to PLS-PM. These criteria allow researchers to guide their model selection efforts in the direction of predictive power, especially in exploratory settings—that is, with an evolving theory base, and under a set of competing models and hypotheses. Each of the model selection criteria described in this study is aimed at selecting the model with the highest out-of-sample prediction power by penalizing model complexity while rewarding model fit. While the information theoretic model selection criteria have a solid standing in the econometrics field (from which PLS-PM originated; Wold (1974)), this is the first study that considers them for model selection in a PLS-PM context. Using a Monte Carlo study, we analyze and compare the performance of the criteria in selecting the best predictive model under various conditions of sample size, effect size, loading patterns, and data distribution.

2. Information theoretic model selection criteria

Model selection criteria that optimize out-of-sample prediction must strike a balance between fitting the particular sample while not over-fitting that sample, so that the model generalizes beyond the particular sample. Achieving this goal is commonly done by combining a measure of model fit with a penalty for model complexity. In the case of linear regression models, one such metric is the adjusted R^2, which includes a penalty proportional to the number of predictors (k) in the model:

$$Adjusted\ R^2 = 1 - [(1-R^2)\ ((n-1)/(n-k-1)]$$

However, the adjusted R^2 lacks formal justification and is not considered a good predictive power metric (Berk, 2008). An alternative specifically designed for predictive purposes is the Final Prediction Error (FPE; Burnham and Anderson, 1998):

$$FPE = \frac{SSE_k}{MSE} + k\lambda_{n,k}$$

Where, SSE_k is the sum of squared errors from a model using k predictors, MSE is the mean squared error using the saturated model with all $p>k$ predictors, and $\lambda_{n,k}$ is a penalty parameter for a model with k predictors and n observations. Two main metrics grounded in information theory emerged from the FPE: Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978), with variations for small samples (AIC_U, AIC_C, Cp, GM, HQ, and HQ_C; McQuarrie and Tsai, 1998):

$$AIC = \log(SSE/n) + \frac{2k}{n}$$

$$BIC = \log(SSE/n) + \frac{k\log(n)}{n}$$

AIC and BIC represent two streams of model selection criteria, which differ fundamentally in their conceptual underpinnings and assumptions. Most importantly, BIC assumes that one of the models in the consideration set is the underlying data generating model and is designed to select the model most likely (in the Bayesian sense) to coincide with the underlying model. In contrast, AIC does not assume that the underlying data generating model is among the set of models under consideration. Instead, AIC is designed to estimate the relative amount of information lost (using the Kullback-Leibler divergence measure between distributions) when a given model estimated from data is compared to a "true" but unknown data generating process.

AIC's strength as a model selection criterion in terms of predictive power has been shown empirically as well as theoretically (Burnham and Anderson, 1998). For example, Stone (1977) showed that the AIC and leave-one-out cross-validation are asymptotically equivalent. One disadvantage of AIC is that it is asymptotically inconsistent, in the sense that if the set of models includes the "true" model (as in the case of a simulation), then the probability of selecting the correct model does not converge to one as the sample size approaches infinity (Shao, 1993). On the contrary, BIC is consistent and, at the same time, puts a heavier penalty than AIC on model complexity.[1] BIC is also related to cross-validation and was shown to be asymptotically equivalent to leave–ν-out cross-validation, where $\nu = n\,(1 - 1/(\log(n) - 1))$.

In light of the criteria's differences, there is no general agreement whether AIC or BIC should be given preference in empirical applications (Shi and Tsai, 2002). Bearing this in mind, this paper considers AIC, BIC and related metrics in the context of out-of-sample prediction in PLS-PM. Prior research has examined the efficacy of these model selection criteria under various conditions in different methodological context such as mixtures of normal distributions (e.g., Biernacki et al., 2000; Bozdogan, 1994; Celeux et al., 1996), mixture regression models (e.g., Andrews and Currim, 2003a; Hawkins et al., 2001; Becker et al., 2014), and mixture logit models (e.g., Andrews and Currim, 2003b). However, the behavior of these criteria for predictive model selection under various model and data conditions is unknown in the context of PLS-PM with the specific goal of out-of-sample prediction. Against this background, we are currently running a Monte Carlo simulation study to explore their performance in the context of PLS-PM.

3. Monte Carlo simulation study

3.1. Study design

The Monte Carlo study analyzes the out-of-sample predictive power of standard model evaluation criteria (i.e, R^2, adjusted-R^2, Q^2, GoF) and the following information theoretic model selection criteria: AIC, AICc, AICu, BIC, Cp, FPE, GM, HQ, and HQc.
The following experimental conditions will be manipulated:

- Six conditions of sample size (50, 100, 150, 200, 250, and 500).
- Five conditions of varying effect size on a structural path (0.1, 0.2, 0.3, 0.4, and 0.5).

[1] The BIC penalty on the number of predictors *k* increases with sample size *n*, unlike in AIC

- Four data distributions (normal, chi-squared distributed with *df* =3, *t* distributed with *df* =5, and uniform) to reflect normal, positively-skewed, heavy-tailed, and uniform distributions, respectively.
- Four factor loading patterns with different levels of average variance extracted (AVE):
    - Higher AVE & homogenous loadings: (0.9, 0.9, 0.9),
    - Higher AVE & heterogeneous loadings: (0.9, 0.8, 0.7),
    - Lower AVE & homogenous loadings: (0.7, 0.7, 0.7), and
    - Lower AVE & heterogeneous loadings: (0.5, 0.6, 0.7)

In addition to the training data, a holdout set (n=1,000) will be created for each experimental condition to mimic the population that the training sample originates from. The dependent variable of interest is a binary variable that will assume the value 1 if the model selection criteria select the model with the highest out-of-sample predictive accuracy (measured in terms of the RMSE), 0 otherwise.

3.2. Expected results

In light of prior research on the efficacy of information theoretic model selection criteria vis-à-vis their in-sample prediction counterparts, we expect the criteria to clearly outperform the R^2, adjusted R^2, and GoF metrics in terms of out-of-sample prediction. At the same time, we expect Q^2 to perform better than R^2 based criteria. Also, we expect some variation in the information criteria's performance, depending on the factor level constellations. For example, we expect that AICc performs favorably for small sample sizes compared to AIC. In addition, BIC is expected to perform well because the set of models include the data generating model. We are in the midst of running the simulation study and expect to present results at the conference.

4.0. REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199-213). New York: Springer.

Andrews, R. L., & Currim, I. S. (2003a). Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing*, *20*(4), 315-321.

Andrews, R. L., & Currim, I. S. (2003b). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, *40*(2), 235-243.

Becker, J. M., Rai, A., & Rigdon, E. (2013). Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance. *34th International Conference on Information Systems*, Milan, Italy.

Becker, J. M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2014). How collinearity affects mixture regression results. *Marketing Letters*, forthcoming.

Bentler, P. M., & Mooijaart, A. B. (1989). Choice of structural model via parsimony: A rationale based on precision. Psychological Bulletin, *106*(2), 315.

Berk, R. (2008). *Statistical learning from regression perspective*. Springer, New York.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719-725.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new

informational measure of complexity. In H. Bozdogan (Ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* Volume 2 (pp. 69-113). Boston: Kluwer Academic Publishers.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. Springer.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195-212.

Evermann, J., & Tate, M. (2014). Comparing out-of-sample predictive ability of PLS, covariance, and regression models. *35th International Conference on Information Systems*, Auckland, Australia.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1-35.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012a). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, *40*(3), 414-433.

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012b). The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5), 320-340.

Hawkins, D. S., Allen, D. M., & Stromberg, A. J. (2001). Determining the number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, *38*(1), 15-48.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science*, *55*(1), 1-34.

Lee, L., Petter, S., Fayard, D., & Robinson, S. (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, *12*(4), 305-328.

McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and Time Series Model Selection* (Vol. 43). Singapore: World Scientific.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190-204.

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-SEM in MIS quarterly. *MIS Quarterly*, *36*(1), iii-xiv.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, *6*(2), 461–464.

Shao J. (1993). Linear model selection by cross-validation. *Journal of American Statistical Association*, *88*(422), 486-494.

Shi, P. and Tsai C.L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society Series B*, *64*(2), 237–252.

Shmueli, G. and Koppius, O. (2011). Predictive Analytics in information systems research. *MIS Quarterly*, *35*(3), 553-572.

Stone M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B. 39*(1), 44–47.

Wold, H. (1974). Causal flows with latent variables: partings of the ways in the light of NIPALS modelling. *European Economic Review*, *5*(1), 67-86.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 47-74). New York.