Comparing Out-of-Sample Predictive Ability of PLS, Covariance, and Regression Models

Completed Research Paper

Joerg Evermann
Memorial University of Newfoundland
St. John's, Canada
jevermann@mun.ca

Mary Tate
Victoria University of Wellington
Wellington, New Zealand
mary.tate@vuw.ac.nz

Abstract

Partial Least Squares Path Modelling (PLSPM) is a popular technique for estimating structural equation models in the social sciences, and is frequently presented as an alternative to covariance-based analysis as being especially suited for predictive modeling. While existing research on PLSPM has focused on its use in causal-explanatory modeling, this paper follows two recent papers at ICIS 2012 and 2013 in examining how PLSPM performs when used for predictive purposes. Additionally, as a predictive technique, we compare PLSPM to traditional regression methods that are widely used for predictive modelling in other disciplines. Specifically, we employ out-of-sample k-fold cross-validation to compare PLSPM to covariance-SEM and a range of a-theoretical regression techniques in a simulation study. Our results show that PLSPM offers advantages over covariance-SEM and other prediction methods.

Keywords: Research Methods, Statistical methods, Structural equation modeling (SEM), Predictive modeling, Partial Least Squares, Covariance Analysis, Regression

Introduction

Explanation and prediction are two main purposes of theories and statistical methods (Gregor, 2006). Explanation is concerned with the identification of causal mechanisms underlying a phenomenon. On the statistical level, explanation is primarily concerned with *testing* the faithful representation of causal mechanisms by the statistical model and the estimation of true population parameter values from samples. Prediction is the ability to predict values for individual cases.

Quantitative research in Information Systems (IS) research has been dominated by causal-explanatory statistical modeling at the expense of a focus on prediction (Shmueli and Koppius, 2011). The IS discipline has traditionally been described as "data poor" because of the difficult and costly way to collect data from companies, managers, and other IT professionals and this has been argued to hold back the generation of theories in the field (Lyytinen, 2009). The advent of "big data" has changed this. Modern IT-based organizations, not only analytics leaders such as Facebook, Google, Amazon and Walmart, but also smaller and less prominent companies, are generating petabytes of data that record billions of digital transactions annually (Davenport, 2006). Carrying out data-driven, predictive modelling on these large datasets has the opportunity to generate fresh insights and drive new theorizing (Shmueli and Koppius, 2011).

Quantitative research in IS frequently uses structural equation modeling, allowing researchers to represent latent constructs, observations, and their relationship in a single statistical model. To estimate such models, researchers typically make use of either covariance-based estimators or the partial least squares path modeling (PLSPM) method. Covariance analysis estimates the model by minimizing the difference between the model-implied and the observed covariance matrices. Because it provides a

statistical test of the fit of data to model, it is typically associated with explanatory modeling. In contrast, the partial least squares path modeling[1] (PLSPM) technique treats the latent constructs as weighted composites of the corresponding observed variables and estimates the composite model using multiple regression. PLSPM is often argued to be preferable to covariance analysis for prediction (Hair, Ringle, and Sarstedt, 2011; Reinartz, Haenlein, and Henseler, 2009).

While in the IS field, perhaps due to its traditional reliance of structural equation models, the method that is strongly associated with prediction is PLSPM, predictive modeling in other research areas (Hastie et al., 2009) uses entirely a-theoretical prediction methods, such as various forms of multiple regression, canonical regression, principal component regression, etc. From the applied researcher's perspective, the main difference between these and structural equation models is that only a very simple model is assumed, omitting elements such as latent variables, measurement error terms, mediation and non-recursive models that are key features of structural equation models. Instead, these techniques deal only with observed predictors and observed outcome variables and typically posit simple linear regression relationships.

Numerous research papers in the past 10 years have focused on evaluating and comparing covariance analysis and the PLSPM technique. However, almost all of them have focused on parameter accuracy and statistical power. These are key issues in testing models and making inferences about population parameters, but are not important to predictive modeling. In contrast, despite the oft-repeated claims about the advantage of PLSPM for predictive modeling (e.g. Ringle et al., 2012; Hair et al., 2011; Hair et al., 2012), there are few studies that have systematically tested these claims. Evermann and Tate (2012) examined predictive ability for reflective factor models using both PLSPM and covariance SEM. Prediction from PLS estimated models, judged by the Q^2 metric based on blindfolded data values, was superior to estimation from covariance SEM estimated models. However, their use of reflective exogenous constructs in the factor models precluded the use of out-of-sample evaluation of the predictive ability through k-fold cross-validation, as is typical in the predictive analytics literature (Hastie et al., 2009). Becker et al. (2013) examined the predictive ability of PLSPM estimated models with formative/composite constructs. While Becker et al. (2013) use of out-of-sample evaluations, they do not focus on the prediction or recoverability of individual scores, but on the R^2 of the regression of the composite formative construct. Further, they use an unidentified model and thus cannot compare PLSPM with covariance estimation.

While a comparison between PLSPM and covariance models for prediction is important, claims about advantages of PLSPM for predictive modeling must also be evaluated in light of well-established a-theoretical predictive techniques: After all, when a researcher claims prediction as her primary aim, a complex structural equation model is not necessary and may not be optimal (McDonald, 1996).

In this paper, we build on the two prior studies by Evermann and Tate (2012) and Becker et al. (2013) to further compare the predictive abilities of covariance and PLSPM models in different situations and for different types of models. Additionally, we also compare these methods to simpler, a-theoretical regression models, such as multiple linear regression, canonical regression, principal component regression etc. The contribution of this study is in the recommendations of which techniques to use, based on a simulation study, to IS researchers that wish to use predictive modeling. While we discuss some expectations with respect to the performance of the different methods, we do not wish to test specific hypotheses but instead explore the performance of different methods across a range of conditions.

The remainder of the paper is structured as follows. We first discuss the nature of prediction and its relationship to causal modeling. We then discuss prediction from structural equation models, followed by a brief introduction to regression-based models. The following sections then present the simulation study,

[1] Although both PLS path modeling and PLS regression are originally developed by Herman Wold, they should not be confused. PLS regression is closer to canonical regression or principal component regression (Hastie et al., 2009, Martens and Næs 1991) and does not estimate path models with latent or composite variables and simultaneous equations (although composites that may be interpreted as proxies for latent variables are created in the process). A good overview of the historical development of PLSPM, with references for further information, is provided in Appendix A of (Sanchez, 2013).

its design, results, and recommendations based on them. This is followed by a discussion and conclusion with further recommendations for researchers.

Causal and Predictive Models

Structural equation models are statistical models that comprise a complex set of regression relationships between observed variables and unobserved variables. These models allow researchers to represent their theories about a research phenomenon. In other words, structural equation models are intended to reflect causal mechanisms. Hence, they are typically associated with causal-explanatory modeling and theory testing, as is evident in most of their use in the IS discipline with a focus on model testing and parameter inference.

While predictive models may be based on causal mechanisms, they need not necessarily be (Gregor, 2006). Martens and Næs (1991; pg. 61f; emphasis in original) write that "We should always *try* to understand what the calibration[2] data tell us. But in multivariate calibration we do not *have to* understand everything *before* we start calibrating! … So the form of the calibration model inside the computer does not necessarily have to reflect a causal explanation of how the … data are generated". Thus, on the statistical level, predictive models may be developed in an exploratory and data-driven way (Shmueli and Koppius, 2011). In fact, Wold (1982a) positioned PLS path modeling to allow for "a dialog with the computer". The aim is not to test whether models accurately represent the causal mechanisms, but instead to identify the best way to predict observations for specific cases that are similar to those in the calibration sample (Shmueli and Koppius, 2011; Martens and Næs, 1991, Wold, 1982).

Thus, in the context of IS research with structural equation models, at one extreme is the causal-explanatory modeling that has traditionally been dominated by covariance-based methods, as they provider richer modelling mechanisms, better parameter estimation, can handle endogeneity to ensure unbiasedness of estimates (Antonakis et al., 2010), and provide tests of model fit (Rönkko and Evermann, 2013). Typically, applications of covariance-based methods make no mention of predictive aims.

At the other extreme are purely predictive models that do away with the structure in structural equation models; in fact, they need not even be regression models (Hastie et al., 2009). Here, there are no requirements to capture the causal mechanisms at all and one finds simple regression-based techniques such as multiple linear regression, canonical regression, or principal component regression. Here, one can also find non-regression based techniques such as k-nearest-neighbors or neural networks (Hastie et al., 2009). These technique only require a researcher to specify observed predictors and predicted variables and do not require the specification of a complex structural equation models representing causal mechanisms.

Thus, at the extreme positions, prediction and explanation have very different aims, employ different techniques and use different models. However, we believe that the causal and predictive modeling do not form a dichotomy but that there is a middle-ground between the two extreme positions. While representing the causal, data generating structures may not be important for the predictive aims of researchers, it can aid in the plausible substantial interpretation of a predictive model, which is often important for pragmatic reasons (Freitas, 2013; Davenport, 2013; Huysmans et al., 2011). For example, a predictive model may be easier to accept by decision makers and other stakeholders when it can be plausibly interpreted. Further, it may be easier to determine the prediction boundaries, i.e. determine under what situations the model will hold and under what situations the model will break, when a plausible substantive interpretation is available. Users of predictive models have more trust in its results, especially for unexpected or counterintuitive predictions, when there is a plausible interpretation available. In contrast to explanatory modeling, the plausible interpretations in this context do not entail a rigorous formal statistical testing of all posited relationships and model constraints as in causal-explanatory modeling.

[2] Martens and Næs (1991) use the term "calibration" in the context of predictive modelling: "to use empirical data and prior knowledge for determining to predict unknown quantitative information from available measurements" (pg. 2).

PLS path modeling was developed to occupy this middle ground and to straddle the traditional divide between causal-explanatory and predictive modeling at the extremes. It aims to maintain interpretability while engaging in predictive modeling. In fact, Herman Wold, who originally developed PLSPM, clearly and explicitly positioned it as a method for prediction (Dijkstra, 1983, 2010; Wold, 1982a), de-emphasizing the importance of statistical tests and inference to population parameters. Lohmöller (1989, pg. 72f) writes about PLS path modeling that "predictor specification is a shortcut term for the type of model building where the investigator

- Starts with the purpose of prediction
- Sets up a system of relations … where the structure of the relations must be founded in the substance of the matter, and the predictive purpose should not jeopardize a structural causal interpretation of the relation.
- … The contrast between predictive vs. structural/causal is not absolute. … For simple models both aspects come at the same time; for complex models there is a parting of the ways."

The emphasis on prediction by PLSPM developers has been picked up by applied researchers in the IS and other management disciplines. Ringle et al. (2012) report that 15% of PLSPM studies in MISQ and almost a quarter of PLSPM studies in other leading management journals claim to focus on prediction. However, few of these researchers act accordingly: Most studies focus their analysis and reporting on parameter estimates and their significance, measurement validity, and causal structures, typically for a single, theory-derived model. Few, if any, studies suggest why prediction rather than causal explanation is of importance in the research context, none perform out-of-sample predictive ability evaluation, and none perform a data-driven exploration of multiple models.

Predicting from Structural Equation Models

In the context of predictive modeling, structural equation models can present unique challenges. Traditionally, in predictive modelling, the values for predicted variables for a new case are predicted from the predictor variables for that case when applied to the estimated ("calibrated") statistical model. However, many structural equation models in the IS discipline are specified in what is termed fully reflective mode. In these models, *all* observed variables are dependent ("endogenous") variables in the model. Only some of the unobserved latent variables are exogenous. Hence, there are no observed predictor variables from which values can be predicted for a new case (Evermann and Tate, 2012). Similarly, when a model is specified in a purely formative way, as was done by Becker et al. (2013), there are no predicted observed variables; the only predicted variables are unobserved, latent variables. Hence, it is difficult to even speak of true prediction also in these cases (Evermann and Tate, 2012).

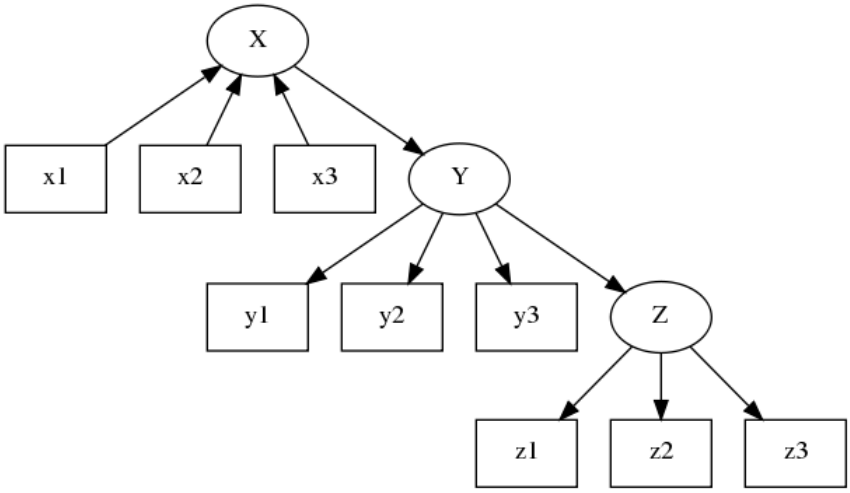

Figure 1: Exogenous formative construct with manifest predictor variables

One can of course simply disregard the directionality of relationships in the structural equation model; after all, it is not important for predictive modeling. This amounts to picking any set of variables and deciding that some are predictors and some are predicted. One can then proceed with evaluating the predictive ability of different statistical models and estimation techniques (at least the a-theoretical ones), but this runs the danger that the exercise is pointless with respect to its applicability in the domain of interest: These types of models suggest that in the substantive domain of interest, there may not be easily identifiable observed predictors or observed predicted variables.

Two previous studies on prediction from structural equation models have dealt with this in different ways. Evermann and Tate (2012) evaluated purely reflective models using a *within-sample* technique and made the assumption that observed variables linked to exogenous latent variables are predictors and that observed variables linked to endogenous latent variables are predicted variables. However, this assumption is, as we pointed out, more or less arbitrary. In contrast, Becker et al. (2013) did not predict individual scores on specific variables from a set of observed scores, but instead evaluated the R^2 of the regression of an endogenous latent variable.

In summary, only structural equation models with a specific form have clear predictors and predicted observed variables. Specifically, this requires that all exogenous latent variables are specified formatively, whereas all endogenous latent variables are specified reflectively (Figure 1). In the example model in Figure 1, it is clear that x1, x2, and x3 are predictors that can be used to predict values for y1 to y3 and z1 to z3.

In summary, while only specific forms of structural equation models allow one to speak of predictive models in the research context that we consider here, where the models should bear some resemblance to the data-generating real world causal mechanism for plausible interpretability, researchers may opt to work with a model that is not of this form when the need to model causal structures dictates this. If that is the case, prediction of the form discussed in this paper, simply cannot be done.

Regression Methods for Prediction

When, as in the example model in Figure 1, there is a clear set of observed predictors and predicted variables in the structural equation model, it is also fair to compare the prediction from such a structural equation model (whether estimated with covariance techniques or PLSPM) to a-theoretical prediction. After all, if the aim of a researcher is prediction, this need not involve a structural equation model at all. While there is a wide range of such methods (Hastie et al., 2009) we focus on regression-based methods as they are closest to the structural equation methods.

Hence, next to covariance-SEM and PLS path modeling, we also compare the performance of multiple linear regression models (LM), principal components regression (PCR), PLS regression, and canonical correlation analysis (CCA) (Hastie et al., 2009; Martens and Næs, 1991). PCR constructs principal components for the independent variables and then estimates regression coefficients between these and the dependent variables. Typically used for data reduction or to avoid multi-collinearity in regression, the model should be equivalent in its predictive abilities to the linear model when all principal components are used for prediction. Similar to PCR, PLS regression also forms orthogonal components of the independent variables, but uses both the dependent and independent variables in their construction, whereas principal components are formed using only the information in the independent variables. Again, the main motivation is data reduction or to avoid multi-collinearity in regression and, just like with PCR, in the limiting case when all components are included in the model, the predictive ability should be equivalent to that of the multivariate linear model. Finally, canonical correlation analysis forms orthogonal composites ("canonical variates") for both sets of variables and estimates regressions between the canonical variates. For m independent and n dependent variables, there exist $\min(m,n)$ matching canonical variates and thus at most $\min(m,n)$ regression coefficients for the prediction.

Both covariance-SEM and PLS path modeling can impose considerable constraints on the predictive model by introducing mediating variables in a non-saturated inner/structural model. Thus, we expect the predictive abilities to be worse than for the simple regression models which impose no such constraints. McDonald (1996) writes with respect to PLSPM prediction that "a path model is … generally subobtimally predictive" and that "if the object of the analysis were to predict the response variables, … we cannot do better than to use a multivariate regression … or the corresponding canonical variate analysis." (pg. 266)

The canonical variates that are estimated in CCA can be interpreted as mediating component variables, albeit data-derived ones rather than theoretically motivated ones as for PLS path modeling[3]. Thus, CCA essentially forms a three stage component path model: from manifest independent variables to their canonical variates, to the canonical variates of the manifest dependent variables, to the manifest dependent variables themselves. However, in contrast to the path models examined here, CCA constructs more variates than there are latent variables in our models. Similarly, the components constructed by PCR and PLS regression can be viewed as data-derived mediating variables. In contrast to CCA, PCR and PLS regression lead to a two-stage path model: from manifest independent variables to the components, to the manifest dependent variables. PCR and PLS regression impose fewer constraints than CCA but more than the LM, as only one set of components mediates between manifest independent and dependent variables. Finally, in the language of path models, the inner models constructed by CCA, PCR and PLS techniques are saturated models.

In summary, we expect the linear regression model to offer the best performance, followed by PCR and PLS regression, followed by CCA, and followed by PLS path modeling or covariance-SEM.

Evaluating Predictive Models using Out-of-Sample Cross-Validation

A widely accepted method to evaluate predictive models is to split a sample in "folds". One part ("fold") of a sample is then used to estimate ("calibrate") the model, the other part of the sample is then used to test the predictive ability of the model: The values for predicted variables are computed from the predictors in the test part of the sample and then compared to their true values in the test part of the sample. To make this procedure more robust to sample variations, in practice a sample is randomly split into *k* folds and the process is called k-fold cross-validation.

Specifically, in k-fold cross-validation, a sample is split randomly into k sub-samples ("folds") of equal size. The following procedure is then repeated k times: Select $k-1$ subsamples as the "training" sample and estimate the model parameters using these observations. In practice, 10-fold cross-validation is "recommended as a good compromise" (Hastie et al. 2009, pg. 243). Using the manifest predictor variables of the remaining "testing" sub-sample and the estimated model parameters, predict the values of the dependent variables. Estimate the mean error of prediction using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), defined as follows:

$$MAE = mean \mid Y_{pred} - Y_{true} \mid$$

$$MSE = mean \left(Y_{pred} - Y_{true} \right)^2$$

$$RMSE = \sqrt{MSE}$$

Study Design, Data Generation, and Model Estimation

To compare the predictive ability of different models and statistical methods, we use a simulation study. A simulation study is a controlled experiment in which observations are generated from a model with parameters fixed by, and known to, the researcher. The different algorithms (covariance SEM, PLSPM, various forms of regression) are then used to estimate the model parameters from the simulated data. The advantage of simulation studies is that different conditions of sample size, numbers of indicators for each latent variable and loadings can be examined. In effect, it constitutes a randomized experiment. As such, it emphasizes internal validity over external validity (i.e. realism).

Models for simulation studies should be representative of those found in the substantive literature (Paxton et al., 2001). PLS models in highly-ranked IS and marketing journals have a median number of 7 to 9 latent variables with 9 to 11 structural relations (Ringle et al., 2012). Another study on the use of PLS in the marketing literature reported a median of 7 latent variables with a median of 8 structural relations

[3] It is this realization that gave Herman Wold the impetus to develop PLS for path models: "I realized that principal components and canonical correlations can be interpreted as path models with one and two latent variables, respectively" (Wold, 1982b as quoted in Sanchez, 2013)

(Hair et al., 2012). Both Ringle et al. (2012) and Hair et al. (2012) report a median of 3.5 indicators for each reflective construct. Based on these reports, we examine the models shown in Figures 2 through 4 (for reasons of space, indicators are not shown). These models are the same as those examined by Evermann and Tate (2012) and were chosen in the interest of comparability of the studies. While models 1 and 2 are relatively simple models, model 3 matches the typical characteristics of PLS models in the literature quite well. Together, they cover model complexity from low to typical.

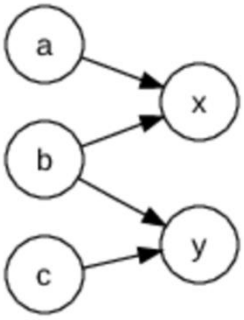

Figure 2: Model 1

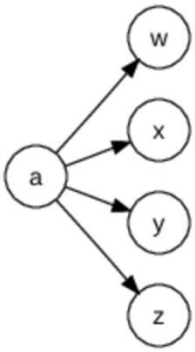

Figure 3: Model 2

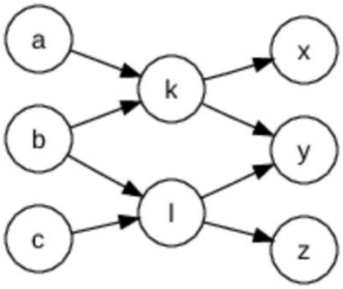

Figure 4: Model 3

Indicators for all exogenous constructs (constructs "a," "b," and "c" in models 1 and 3, and construct "a" in model 2) are formatively specified, while indicators for endogenous constructs (constructs "x" and "y" in model 1, constructs "w" through "z" in model 2, and constructs "k," "l," "x," "y," and "z" in model 3) are reflectively specified. Sample size, the number of indicators, indicator loadings, structural effect sizes and measurement specification (formative or reflective) are important design variables for PLS simulation studies (Aguirre-Ureta and Marakas, 2008; Becker et al., 2013; Goodhue et al., 2007; Goodhue et al., 2012; Reinartz et al., 2009). Table 1 shows a summary of the experimental factors in this study.

Design factor		**Levels**
s	Sample size	100, 250, 750
i	Number of indicators	3, 5, 7
l	Indicator loadings (non-std.)	1
b	Effect sizes (β, γ)	.75
c	Indicator correlation	0, 0.1, 0.4
e	Error variance formative construct	0, 0.1, 0.4
—	Sampling distribution	Normal
—	Response type	Continuous
—	Missing values	None
—	Measurement model	Mixed (formative and reflective)

Table 1: Study Design Factors

Following Evermann and Tate (2012), we used three different sample sizes (100, 250, and 750) and three different indicator counts for each construct (3, 5, and 7). To keep the study design simple and the computational requirements manageable, we used loadings of 1 and an error variance of 0.1 for the reflective indicators. For the same reasons, we also used only a single effect size (0.75) for the inner/structural model relationships.

As there is little information in the literature on the typical correlations of formative indicators of the same construct, we examined the ideal situation (correlation = 0), a relatively bad situation (correlation = 0.4) and a moderate situation (correlation = 0.1). Similarly, guidelines suggest that formative indicators should capture all of the construct (i.e. constitute the complete set of predictors), thus leaving no error (residual) variance for the formative construct. We examined the ideal situation (error variance = 0), a relatively bad situation (error variance = 0.4) and a moderate situation (error variance = 0.1).

In summary, our study design yields 3 x 3 x 3 x 3 = 81 conditions for each of the three models. For each of these experimental conditions, we generated and estimated 200 samples.

We estimated each sample using the following estimation and prediction methods:

- LM
- PCR
- CCA
- PLS regression, using the following algorithms
    - Canonical Powered PLS (CPPLS) (Indahl, et al., 2009)
    - Kernel PLS algorithm (KPLS) (Lindgren et al., 2005)
    - Wide kernel PLS algorithm (WKPLS) (Mevik et al. 2011)
    - Orthogonal scores algorithm (OSCORESPLS) (Melvik et al. 2011)
    - SIMPLS (de Jong, 1993)
- PLS path modeling
- Covariance-SEM

PLS path modeling offers the choice of using mode A or mode B estimation of the composite weights. While these are different ways to construct composites during the outer estimation phase of the PLS algorithm and do not affect the subsequent estimation of model parameters (Becker et al., 2013; Rönkko and Evermann, 2013), we follow established research practice and use mode A for reflective constructs and mode B for formative constructs. While Becker et al. (2013) observed better performance for mode A estimation for small sample sizes, their results are not directly comparable to our study as they pertain to the R^2 of the endogenous composite, whereas we are interested in the RMSE for the dependent observed variables.

For prediction from PCR, CCA and PLS regression, we used the maximal number of components or canonical variates. For each sample and estimation/prediction method, we applied 10-fold blindfolding. For the outcome measures, we report the mean RMSE values over all folds and samples[4].

Results and Recommendations

Similar to the findings by Evermann and Tate (2012), we have found no model-dependent results, so that the following presentation and discussion of results hold for all models examined in the study. The results for model 3 of this study are presented in the appendix. Figures 5 and 6 show the data for our most complex model 3 presented in two different ways. Figure 5 shows the RMSE for different combinations of sample size, number of indicators, error variance and estimation/prediction method.

[4] Because these are strongly monotonous functions, the relative performance of the estimation/prediction methods w.r.t MAE and MSE are maintained, so that reporting any one of the three is sufficient for our analysis.

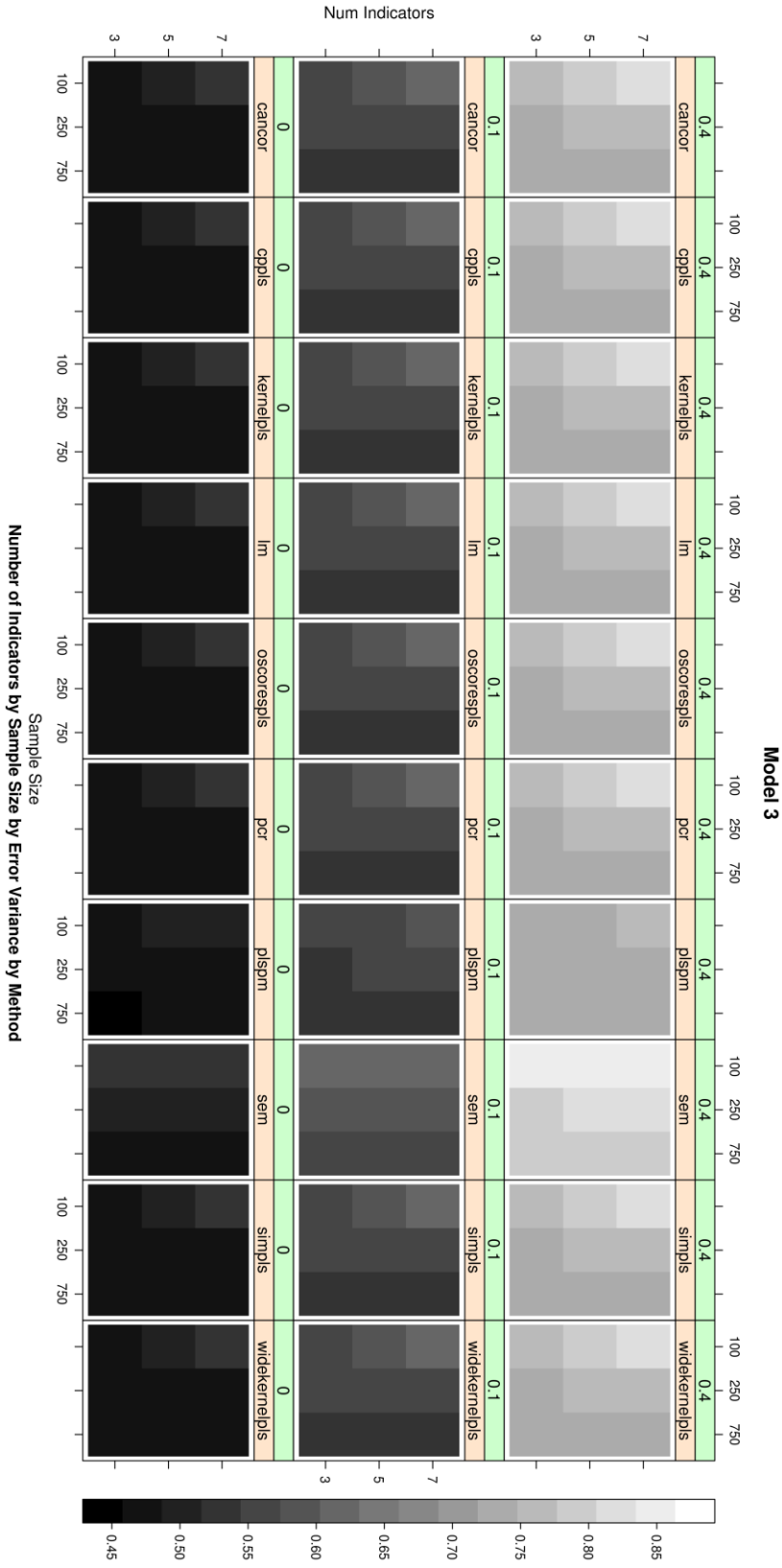

Figure 5: RMSE for model 3 by sample size, number of indicators, error variance and method

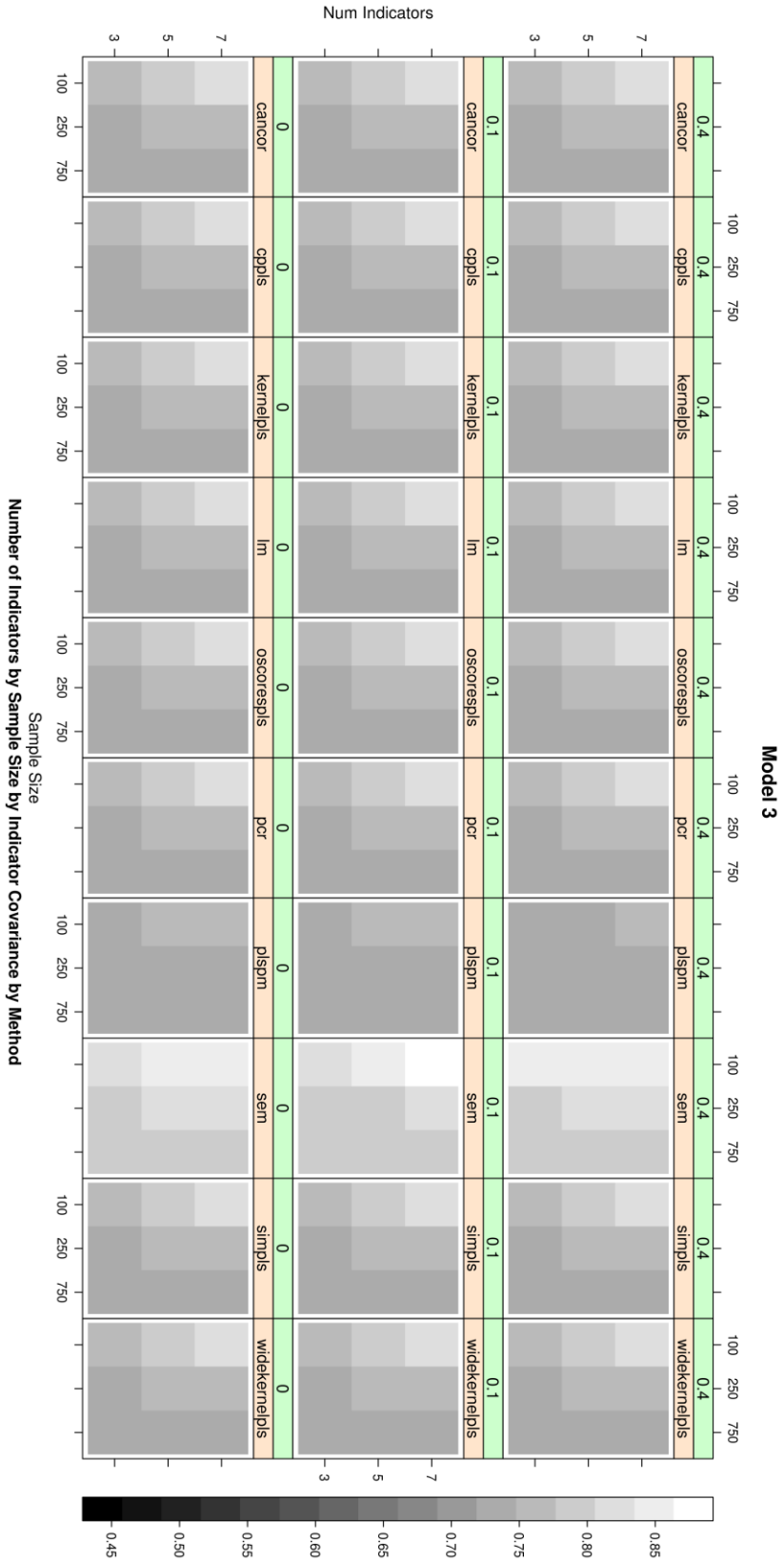

Figure 6: RMSE for model 3 by sample size, number of indicators, indicator covariance and method

We can see that the RMSE increases with increasing error variance. This is expected, as any additional random error in the prediction path between independent and dependent variables will negatively impact the predictive abilities of the model no matter what estimation or prediction method is chosen. We also see that better prediction is achieved with fewer indicators and higher sample size. This is especially obvious in Figure 5 when the error variance is high, but present in all other conditions and for all prediction methods. Again, this is not surprising as a larger sample size means more accurate estimations of model parameters (smaller standard errors) and fewer variables mean that a model with the same number of structural parameters or constraints has to predict fewer dependent values.

From Figure 6 we can see that there are no differences in RMSE due to different levels of indicator covariance. This is as expected, because multi-collinearity affects only the parameter standard errors, not their point estimates, and it is only the latter that are used for prediction[5]. Again we see the effect of sample size and number of indicators, as from Figure 5. We also note the difference between PLS path modeling and covariance-SEM. As in Figure 5, PLS path modeling is marginally better than the various regression methods, whereas covariance-SEM is significantly worse in all conditions.

The two key results in this study are visible in Figures 5 and 6, and highlighted in Table 3 in the appendix. *First, the performance of covariance-SEM was worse than that of PLS path modeling and that of the various regression methods in all experimental conditions. Second, and counter to our expectations, PLS path modeling performs slightly better than even the various regression models in all experimental conditions.*

The various regression methods are essentially equivalent in their predictive abilities, which is due to the use of the full set of canonical variables or composites that are created in the process. If fewer than the full set had been used, we might have seen differences between these prediction methods. However, in this paper we were not interested in dimensionality reduction but only in optimal prediction.

Our results support the claims in the literature that PLS path modeling is indeed a good choice for predictive purposes in structural equation modeling. Based on our results from this study, we make the following recommendations when individual scores are to be predicted from structural equation models:

> *When predicting from structural equation models where all exogenous constructs are formative, researchers should use PLS path modelling for prediction.*

We were surprised by the good performance of the PLS path modeling algorithm compared to the regression methods. While this may be counter-intuitive and is counter to McDonald's (1996) assertion, we suggest that this is due to the fact that the estimated model is identical to the generating model. Thus, the constraints imposed by the path model do not actually affect the model estimates and prediction from the model, whereas the data-driven components formed in PCR, PLS regression of CCA are unlikely to coincide with the formative latent variables in the PLS path model and thus impose effective constraints on the model. Hence, this advantage of PLS path modeling over regression methods may not hold for mis-specified models.

Unfortunately, model misspecification is a condition which is difficult to identify with PLSPM (Evermann and Tate, 2010) and can better be dealt with using covariance techniques (Antonakis et al., 2010; Rönkko and Evermann, 2013). Given the limitations of our present research in this respect, and until future work is done to either support or rule out our conjecture about the correctness requirement of the model, our recommendation must be qualified to recommend a model test prior to prediction:

> *When predicting from structural equation models, researchers should ensure the correctness of the model using covariance techniques and subsequently predict with PLS path modeling.*

[5] However, parameter standard errors enter into the computation of prediction confidence intervals, which have recently been proposed as a measure of pragmatic validity (Lee and Hubona, 2009)

Discussion and Conclusion

The major contribution of this study is an evaluation of the out-of-sample predictive ability of two competing methods for estimating formative structural equation models, PLS path modeling and covariance-SEM, and a comparison to purely predictive, a-theoretical regression models.

This study has shown that PLSPM is a more appropriate choice than covariance SEM when the goal is prediction from structural equation models that contain a clear set of predictors. In contrast to simple regression methods, PLS path modelling allows the researcher to specify a prediction model that maintains the causal interpretability of the model parameters and model structure, as argued by Lohmöller (1989). In line with his statements (quoted earlier), in our understanding of PLS path modelling, the underlying model serves to make the prediction "plausible" and the model "interpretable"; it is not necessarily a proven correct casual model that allows inference to population parameter values.

In fact, we have recommended that researchers use covariance SEM and PLS path modeling in complementary ways, as suggested by Hair et al. (2012). Specifically, covariance SEM can establish model correctness (including an assessment of measurement validity and reliability) to ensure that the subsequent prediction from that model with PLS path modelling remains efficient (due to the correctness of the model) and interpretable or plausible[6].

In such a combined application, covariance SEM will estimate *population parameters*, whereas the PLS estimation will result in *sample parameters* for prediction from a particular sample, with no claim of inference to the population. Model fit, parameter significance tests, validity and reliability assessment, goodness of fit indices, etc. are appropriate for the covariance SEM estimation but not for the PLSPM estimation. In contrast, predictive validity assessments are appropriate for the PLS estimation, but not the covariance estimation, as prediction is sample or application specific. If prediction is the main aim of a study, the fact that a model shows lack of fit by traditional metrics, such as the PLS goodness-of-fit index or the various fit indices for covariance-based estimation, or lack of validity and reliability, is irrelevant. Such a refocusing of PLS path modeling away from the evaluation of factor-based path models and towards prediction reflects Dijkstra's (2010) recent call: "We also propose a new 'back-to-basics' research program, moving away from factor analysis models and returning to the original object of constructing indices that extract information from high-dimensional data in a predictive, useful way. … Cross-validation could settle the choice between various competing specifications" (pg. 23)

Our final recommendation, to researchers who motivate their choice of PLSPM by appealing to predictive aims of their study, is to act accordingly. The theoretical motivation, parameter significance testing, and the assessment of validity and reliability of the measures should take a limited and sub-ordinate role (if it is discussed at all) to the data-driven exploration of predictive models and the presentation of appropriate blindfolding or cross-validation results. Researchers should make an explicit case for how the prediction of specific cases is an appropriate and relevant goal for their study, rather than merely mentioning this in passing. Given the current execution and reporting of many PLSPM based studies in IS research practice, with their emphasis on theoretical motivation, single models, validity and reliability testing, and parameter significance tests, it is difficult to accept the claims of predictive goals that many published studies use to motivate their choice of PLSPM.

In summary, in line with Shmueli and Koppius (2011), we believe that prediction is an under-utilised approach in Information Systems Research, and one that deserves more attention. It is essential that appropriate methodological choices are made for predictive modelling. We contribute to this discussion by examining the relative performance of PLS, covariance based models and regression models under various conditions by showing that PLS path modeling has a place in the methodological toolbox of IS researchers for predictive modelling, and outperforms covariance-based models and even regression models across a range of simulated conditions.

[6] This assumes that the covariance model can be estimated using PLSPM, as covariance estimation provides richer modeling features (e.g. correlated errors across constructs, etc.)

APPENDIX

The appendix contains the RMSE results for model 3 in tabular form. As the relative performance of each prediction method was independent of the model, and due to space limitations, we present complete information only for one model. Each row in the tables represents one of the 81 experimental conditions. The columns CCA to Wkernel PLS represent the RMSE metric of the prediction error. Estimation was performed using the plspm, lavaan, and pls packages for the R statistical system.

Lowest RMSE is in **bold**, highest RMSE is underlined.

Sample Size	Num Ind	Ind Cov	Error Var	CCA	CPPLS	Kernel PLS	LM	OSCORE S PLS	PCR	PLS PM	SEM	SIM PLS	Wkernel PLS
100	3	0	0	0.481	0.481	0.481	0.481	0.481	0.481	**0.4707**	0.5527	0.481	0.481
100	3	0	0.1	0.5671	0.5671	0.5671	0.5671	0.5671	0.5671	**0.5502**	0.6089	0.5671	0.5671
100	3	0	0.4	0.761	0.761	0.761	0.761	0.761	0.761	**0.7319**	0.8212	0.761	0.761
100	3	0.1	0	0.4817	0.4817	0.4817	0.4817	0.4817	0.4817	**0.4703**	0.5402	0.4817	0.4817
100	3	0.1	0.1	0.5672	0.5672	0.5672	0.5672	0.5672	0.5672	**0.5500**	0.6117	0.5672	0.5672
100	3	0.1	0.4	0.7623	0.7623	0.7623	0.7623	0.7623	0.7623	**0.7337**	0.8245	0.7623	0.7623
100	3	0.4	0	0.4823	0.4823	0.4823	0.4823	0.4823	0.4823	**0.4714**	0.5281	0.4823	0.4823
100	3	0.4	0.1	0.5662	0.5662	0.5662	0.5662	0.5662	0.5662	**0.5497**	0.6097	0.5662	0.5662
100	3	0.4	0.4	0.7662	0.7662	0.7662	0.7662	0.7662	0.7662	**0.7364**	0.8351	0.7662	0.7662
100	5	0	0	0.5022	0.5022	0.5022	0.5022	0.5022	0.5022	**0.4876**	0.5389	0.5022	0.5022
100	5	0	0.1	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	**0.5661**	0.6107	0.5897	0.5897
100	5	0	0.4	0.7987	0.7987	0.7987	0.7987	0.7987	0.7987	**0.7526**	0.8477	0.7987	0.7987
100	5	0.1	0	0.5024	0.5024	0.5024	0.5024	0.5024	0.5024	**0.4885**	0.529	0.5024	0.5024
100	5	0.1	0.1	0.5867	0.5867	0.5867	0.5867	0.5867	0.5867	**0.5621**	0.6134	0.5867	0.5867
100	5	0.1	0.4	0.8029	0.8029	0.8029	0.8029	0.8029	0.8029	**0.7555**	0.8495	0.8029	0.8029
100	5	0.4	0	0.5018	0.5018	0.5018	0.5018	0.5018	0.5018	**0.4883**	0.5263	0.5018	0.5018

100	5	0.4	0.1	0.589	0.589	0.589	0.589	0.589	0.589	**0.5638**	0.6144	0.589	0.589
100	5	0.4	0.4	0.792	0.792	0.792	0.792	0.792	0.792	**0.7469**	0.8367	0.792	0.792
100	7	0	0	0.5231	0.5231	0.5231	0.5231	0.5231	0.5231	**0.5025**	0.5603	0.5231	0.5231
100	7	0	0.1	0.6182	0.6182	0.6182	0.6182	0.6182	0.6182	**0.5806**	0.8634	0.6182	0.6182
100	7	0	0.4	0.8252	0.8252	0.8252	0.8252	0.8252	0.8252	**0.7597**	0.8488	0.8252	0.8252
100	7	0.1	0	0.5235	0.5235	0.5235	0.5235	0.5235	0.5235	**0.5026**	0.5358	0.5235	0.5235
100	7	0.1	0.1	0.6154	0.6154	0.6154	0.6154	0.6154	0.6154	**0.5812**	0.622	0.6154	0.6154
100	7	0.1	0.4	0.8288	0.8288	0.8288	0.8288	0.8288	0.8288	**0.7609**	0.8641	0.8288	0.8288
100	7	0.4	0	0.5252	0.5252	0.5252	0.5252	0.5252	0.5252	**0.5059**	0.5293	0.5252	0.5252
100	7	0.4	0.1	0.6166	0.6166	0.6166	0.6166	0.6166	0.6166	**0.5817**	0.6235	0.6166	0.6166
100	7	0.4	0.4	0.8298	0.8298	0.8298	0.8298	0.8298	0.8298	**0.7636**	0.8598	0.8298	0.8298
250	3	0	0	0.4657	0.4657	0.4657	0.4657	0.4657	0.4657	**0.4620**	0.4873	0.4657	0.4657
250	3	0	0.1	0.5464	0.5464	0.5464	0.5464	0.5464	0.5464	**0.5405**	0.5733	0.5464	0.5464
250	3	0	0.4	0.7379	0.7379	0.7379	0.7379	0.7379	0.7379	**0.7274**	0.7978	0.7379	0.7379
250	3	0.1	0	0.4655	0.4655	0.4655	0.4655	0.4655	0.4655	**0.4619**	0.4871	0.4655	0.4655
250	3	0.1	0.1	0.5466	0.5466	0.5466	0.5466	0.5466	0.5466	**0.5404**	0.5723	0.5466	0.5466
250	3	0.1	0.4	0.7393	0.7393	0.7393	0.7393	0.7393	0.7393	**0.7288**	0.7994	0.7393	0.7393
250	3	0.4	0	0.4655	0.4655	0.4655	0.4655	0.4655	0.4655	**0.4620**	0.4868	0.4655	0.4655
250	3	0.4	0.1	0.5477	0.5477	0.5477	0.5477	0.5477	0.5477	**0.5412**	0.5747	0.5477	0.5477
250	3	0.4	0.4	0.7398	0.7398	0.7398	0.7398	0.7398	0.7398	**0.7293**	0.7998	0.7398	0.7398
250	5	0	0	0.472	0.472	0.472	0.472	0.472	0.472	**0.4684**	0.4881	0.472	0.472
250	5	0	0.1	0.5536	0.5536	0.5536	0.5536	0.5536	0.5536	**0.5453**	0.5738	0.5536	0.5536
250	5	0	0.4	0.7515	0.7515	0.7515	0.7515	0.7515	0.7515	**0.736**	0.8075	0.7515	0.7515
250	5	0.1	0	0.4713	0.4713	0.4713	0.4713	0.4713	0.4713	**0.4674**	0.4889	0.4713	0.4713

250	5	0.1	0.1	0.5554	0.5554	0.5554	0.5554	0.5554	0.5554	**0.5476**	0.574	0.5554	0.5554
250	5	0.1	0.4	0.7476	0.7476	0.7476	0.7476	0.7476	0.7476	**0.7322**	0.8025	0.7476	0.7476
250	5	0.4	0	0.4719	0.4719	0.4719	0.4719	0.4719	0.4719	**0.4677**	0.4874	0.4719	0.4719
250	5	0.4	0.1	0.5542	0.5542	0.5542	0.5542	0.5542	0.5542	**0.5469**	0.5745	0.5542	0.5542
250	5	0.4	0.4	0.7511	0.7511	0.7511	0.7511	0.7511	0.7511	**0.7357**	0.8058	0.7511	0.7511
250	7	0	0	0.4797	0.4797	0.4797	0.4797	0.4797	0.4797	**0.4757**	0.4899	0.4797	0.4797
250	7	0	0.1	0.5639	0.5639	0.5639	0.5639	0.5639	0.5639	**0.5535**	0.5793	0.5639	0.5639
250	7	0	0.4	0.7603	0.7603	0.7603	0.7603	0.7603	0.7603	**0.7393**	0.8091	0.7603	0.7603
250	7	0.1	0	0.479	0.479	0.479	0.479	0.479	0.479	**0.4758**	0.4895	0.479	0.479
250	7	0.1	0.1	0.5628	0.5628	0.5628	0.5628	0.5628	0.5628	**0.5534**	0.5782	0.5628	0.5628
250	7	0.1	0.4	0.7594	0.7594	0.7594	0.7594	0.7594	0.7594	**0.7387**	0.8081	0.7594	0.7594
250	7	0.4	0	0.4792	0.4792	0.4792	0.4792	0.4792	0.4792	**0.4745**	0.4886	0.4792	0.4792
250	7	0.4	0.1	0.5627	0.5627	0.5627	0.5627	0.5627	0.5627	**0.5526**	0.5773	0.5627	0.5627
250	7	0.4	0.4	0.7589	0.7589	0.7589	0.7589	0.7589	0.7589	**0.7391**	0.8089	0.7589	0.7589
750	3	0	0	0.4584	0.4584	0.4584	0.4584	0.4584	0.4584	**0.4575**	0.4748	0.4584	0.4584
750	3	0	0.1	0.5378	0.5378	0.5378	0.5378	0.5378	0.5378	**0.5362**	0.5619	0.5378	0.5378
750	3	0	0.4	0.7282	0.7282	0.7282	0.7282	0.7282	0.7282	**0.7249**	0.7906	0.7282	0.7282
750	3	0.1	0	0.458	0.458	0.458	0.458	0.458	0.458	**0.4571**	0.4741	0.458	0.458
750	3	0.1	0.1	0.5388	0.5388	0.5388	0.5388	0.5388	0.5388	**0.5372**	0.5626	0.5388	0.5388
750	3	0.1	0.4	0.7301	0.7301	0.7301	0.7301	0.7301	0.7301	**0.7268**	0.7931	0.7301	0.7301
750	3	0.4	0	0.4582	0.4582	0.4582	0.4582	0.4582	0.4582	**0.4573**	0.4747	0.4582	0.4582
750	3	0.4	0.1	0.5388	0.5388	0.5388	0.5388	0.5388	0.5388	**0.5373**	0.5625	0.5388	0.5388
750	3	0.4	0.4	0.7281	0.7281	0.7281	0.7281	0.7281	0.7281	**0.7248**	0.7905	0.7281	0.7281
750	5	0	0	0.4605	0.4605	0.4605	0.4605	0.4605	0.4605	**0.4595**	0.4751	0.4605	0.4605

750	5	0	0.1	0.5418	0.5418	0.5418	0.5418	0.5418	0.5418	**0.5395**	0.5639	0.5418	0.5418
750	5	0	0.4	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312	**0.7264**	0.7921	0.7312	0.7312
750	5	0.1	0	0.4604	0.4604	0.4604	0.4604	0.4604	0.4604	**0.4595**	0.4749	0.4604	0.4604
750	5	0.1	0.1	0.5413	0.5413	0.5413	0.5413	0.5413	0.5413	**0.5391**	0.5631	0.5413	0.5413
750	5	0.1	0.4	0.7317	0.7317	0.7317	0.7317	0.7317	0.7317	**0.7271**	0.7929	0.7317	0.7317
750	5	0.4	0	0.4602	0.4602	0.4602	0.4602	0.4602	0.4602	**0.4592**	0.4746	0.4602	0.4602
750	5	0.4	0.1	0.5415	0.5415	0.5415	0.5415	0.5415	0.5415	**0.539**	0.5633	0.5415	0.5415
750	5	0.4	0.4	0.7328	0.7328	0.7328	0.7328	0.7328	0.7328	**0.728**	0.7938	0.7328	0.7328
750	7	0	0	0.462	0.462	0.462	0.462	0.462	0.462	**0.4612**	0.4747	0.462	0.462
750	7	0	0.1	0.5436	0.5436	0.5436	0.5436	0.5436	0.5436	**0.541**	0.5641	0.5436	0.5436
750	7	0	0.4	0.7355	0.7355	0.7355	0.7355	0.7355	0.7355	0.7294	0.7952	0.7355	0.7355
750	7	0.1	0	0.4625	0.4625	0.4625	0.4625	0.4625	0.4625	**0.4615**	0.4754	0.4625	0.4625
750	7	0.1	0.1	0.5443	0.5443	0.5443	0.5443	0.5443	0.5443	**0.5415**	0.5646	0.5443	0.5443
750	7	0.1	0.4	0.7353	0.7353	0.7353	0.7353	0.7353	0.7353	**0.7294**	0.795	0.7353	0.7353
750	7	0.4	0	0.4624	0.4624	0.4624	0.4624	0.4624	0.4624	**0.4614**	0.4752	0.4624	0.4624
750	7	0.4	0.1	0.5445	0.5445	0.5445	0.5445	0.5445	0.5445	**0.5417**	0.5648	0.5445	0.5445
750	7	0.4	0.4	0.7352	0.7352	0.7352	0.7352	0.7352	0.7352	**0.7291**	0.7944	0.7352	0.7352

Table A1: 10-fold cross-validation RMSE errors for Model 3

REFERENCES

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. 2010. "On making causal claims: A review and recommendations." *Leadership Quarterly* (21), 1086-1120.

Becker, J.-M., Rai, A., and Rigdon, E. 2013. "Predictive Validity and Formative Measurement in Structural Equation Modeling: Embracing Practical Relevance". *Proceedings of the 32nd International Conference on Information Systems (ICIS),* Milan, Italy.

Davenport, T.H. 2006. "Competing on analytics." *Harvard Business Review* (84:1), 98-107.

Davenport, T.H. 2013. "Keep up with your Quants." *Harvard Business Review* (91), 120-123.

De Jong, S. 1993. "SIMPLS: An alternative approach to partial least squares regression." *Chemometrics and Intelligent Laboratory Systems.* (18), 251-263.

Dijkstra, T. 1983. "Some comments on maximum-likelihood and partial least squares methods." *Journal of Econometrics,* (22), 67-90.

Dijkstra, T. 2010. "Latent variables and indices: Herman Wold's basic design and Partial Least Squares." In E. Esposito Vinzi et al. (eds.) *Handbook of Partial Least Squares.* Springer-Verlag: Berlin, pp. 23-46.

Evermann, J. and Tate, M. 2010. "Testing models or fitting models? Identifying model misspecifications in PLS." *Proceedings of the 31st International Conference on Information Systems (ICIS),* St. Louis, MS.

Evermann, J. and Tate, M. 2012. "Comparing the Predictive Ability of PLS and Covariance Analysis." *Proceedings of the 33rd International Conference on Information Systems (ICIS)*, Orlando, FL.

Freitas, A.A. 2013. "Comprehensible Classification Models – A Position Paper." *ACM SIGKDD Explorations Newsletter,* (15), 1-10.

Goodhue, D., Lewis, W., and Thompson, R. 2007. "Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators." *Information Systems Research,* (18)*,* 211-227.

Goodhue, D., Lewis, W., and Thompson, R. 2012. "Comparing PLS to Regression and LISREL: A Response to Marcoulides, Chin, and Saunders," *MIS Quarterly* (36:3), 703-716.

Gregor, S. 2006. "The nature of theory in information systems." *MIS Quarterly,* (30), 611-642.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Verlag: Berlin, Germany.

Hair, J.F, Ringle C.M. and Sarstedt, M. 2011. "PLS-SEM: Indeed a silver bullet." *Journal of Marketing Theory and Practice* (19)*,* 139-151.

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

Henseler, J., Ringle, C.M. and Sinkovics, R.R. 2009. "The Use of Partial least squares path modeling in International Marketing." *Advances in International Marketing,* (20), 277-319.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. 2011. "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models." *Decision Support Systems,* (51), 141-154.

Indahl, U.G., Hovde Liland, K. and Næs, T. 2009. "Canonical partial least squares – a unified PLS approach to classification and regression problems." *Journal of* Chemometrics, (23), 495-504.

Lee, A.S., and Hubona, G.S. 2009. "A Scientific Basis for Rigor and Relevance in Information Systems Research," *MIS Quarterly,* (33), 237-262.

Lindgren, F., Geladi, P. and Wold, S. 2005. "The kernel algorithm for PLS." *Journal of Chemometrics.* (7), 45-59.

Lohmöller, J.B. 1989. *Latent Variable Path Modeling with Partial Least Squares.* Physica Verlag: Heidelberg, Germany.

Lyytinen, K. 2009. "Data matters in IS theory building." *Journal of the Association for Information Systems* (10:10), 715-720.

Martens, H., Næs, T. 1989. *Multivariate calibration.* Wiley: Chichester, England.

McDonald, R.P. 1996. "Path analysis with composite variables." *Multivariate Behavioral Research,* (31)*,* 239-270.

Mevik, B.-H., Wehrens, R., Hovde Liland, K. 2011. "pls: Partial Least Squares and Principal Component regression". R package version 2.3-0. http://CRAN.R-project.org/package=pls

Reinartz, W., Haenlein, M., and Henseler, J. 2009. "An empirical comparison of the efficacy of covariance-based and variance-based SEM." *International Journal of Research in Marketing,* (26)*,* 332-344.

Ringle, C.M., Sarstedt, M., and Straub, D.W. 2012. "A critical look at the use of PLS-SEM in MIS Quarterly." *MIS Quarterly,* (36)*,* iii-xiv.

Rönkkö, M., Evermann, J. 2013. „A critical examination of common beliefs about Partial Least Squares Path Modeling." *Organizational Research Methods*. Preprint published online 7 March 2013.

Sanchez, G. 2013. *PLS Path Modeling with R*. Online at http://`www.gastonsanchez.com`

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

Wold, H. 1982a. "Soft Modeling: The basic design and some extensions" In: Jöreskog, K, Wold, S. and Wold, H. (eds.) *Systems under indirect observation: Causality, Structure, Prediction*. Elsevier Publishing Company: North Holland.

Wold, H. 1982b. "Models for Knowledge" In: Gani, J. (ed.) *The Making of the Statistician*. Springer Verlag: Heidelberg, Germany.